United Nations ECE/CES/2014/1

Economic and Social Council

Distr.: General
27 January 2014

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-second plenary session
Paris, 9-11 April 2014
Item 5 of the provisional agenda
Work of the High-level Group for the Modernization of Statistical Production and Services

Generic Statistical Business Process Model

Prepared by the High-Level Group for the Modernization of Statistical Production and Services

Summary

This document provides a description of the Generic Statistical Business Process Model and how it relates to other key standards for statistical modernisation. The Generic Statistical Business Process Model describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonised terminology to help statistical organisations to modernise their statistical production processes, as well as to share methods and components. The Generic Statistical Business Process Model can also be used for integrating data and metadata standards, as a template for process documentation, for harmonizing statistical computing infrastructures, and to provide a framework for process quality assessment and improvement.

The document is submitted to the Conference of European Statisticians for information.

GE.14-20381

1420381

Please recycle

Contents

		Page
I.	Introduction	3
	A. Background	3
	B. Main changes from version 4.0	4
II.	The Model	4
	A. Understanding the Generic Statistical Business Process Model	4
	B. The structure	4
	C. Applicability	6
	D. Using the Generic Statistical Business Process Model	6
III.	Relationships with other models and standards	7
	A. Common Statistical Production Architecture	7
	B. Generic Statistical Information Model	7
	C. Generic Longitudinal /Business Process Model	8
IV.	Levels 1 and 2 of the Generic Statistical Business Process Model	9
V.	Descriptions of phases and sub-processes	10
	A. Specify Needs Phase	10
	B. Design Phase	11
	C. Build Phase	13
	D. Collect Phase	15
	E. Process Phase	17
	F. Analyse Phase	19
	G. Disseminate Phase	20
	H. Evaluate Phase	21
VI.	Over-arching processes	22
	A. Quality management	22
	B. Metadata management	24
VII.	Other uses of the Generic Statistical Business Process Model	25
Annex 1 – List of acronyms		27

I. Introduction

1. The Generic Statistical Business Process Model (GSBPM) describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonised terminology to help statistical organisations to modernise their statistical production processes, as well as to share methods and components. GSBPM can also be used for integrating data and metadata standards, as a template for process documentation, for harmonizing statistical computing infrastructures, and to provide a framework for process quality assessment and improvement. These and other purposes for which GSBPM can be used are elaborated further in Section VI. This version of GSBPM is fully aligned with version 1.1 of the Generic Statistical Information Model (GSIM)[1] and provides a basis for the implementation of the Common Statistical Production Architecture (CSPA)[2].

A. Background

2. The Joint UNECE / Eurostat / OECD Work Sessions on Statistical Metadata (METIS) have prepared a Common Metadata Framework (CMF)[3]. Part C of this framework is entitled "Metadata and the Statistical Cycle". This part refers to the phases of the statistical business process (also known as the statistical value chain or statistical cycle) and provides generic terms to describe them. Since November 2013, this work has been taken over by the Modernisation Committee on Standards, under the High-Level Group for the Modernisation of Statistical Production and Services (HLG)[4].

3. During a workshop on the development of Part C of the CMF, held in Vienna in July 2007[5], the participants agreed that the business process model used by Statistics New Zealand would provide a good basis for developing a Generic Statistical Business Process Model. Following several drafts and public consultations, version 4.0 of GSBPM was released in April 2009. It was subsequently widely adopted by the global official statistics community, and formed one of the cornerstones of the HLG vision and strategy for standards-based modernisation.

4. In December 2012, a complementary model, the Generic Statistical Information Model (GSIM) was released (for more information see section III). The work to develop and subsequently implement GSIM resulted in the identification of several possible enhancements to GSBPM. During 2013, the HLG launched a project on "Frameworks and Standards for Statistical Modernisation" which included a broader review of GSBPM and GSIM, to improve consistency between the documentation of the models, and to incorporate feedback based on practical implementations. The current version of GSBPM (version 5.0) is the direct result of this work. Whilst it is considered final at the time of release, it is also expected that future updates may be necessary in the coming years, either to reflect further experiences from implementing the model in practice, or due to the evolution of the nature of statistical production. The reader is therefore invited to check the website www.unece.org/stats/gsbpm to be sure of having the latest version.

---

[1] See: http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

[2] See: http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home

[3] See: http://www.unece.org/stats/cmf/

[4] See: http://www1.unece.org/stat/platform/display/hlgbas

[5] The papers from this Workshop are available at: http://www.unece.org/stats/documents/2007.07.metis.htm

B. Main changes from version 4.0

5. The review of version 4.0 of GSBPM had a clear mandate from the HLG to only introduce changes that had a strong business case and widespread support in the international statistical community. The rationale for this approach was that many organisations had invested heavily in implementing GSBPM. Major changes could imply significant costs, and could ultimately be counter-productive in terms of securing the widespread adoption and use of the model. As a result, there are a limited number of changes to the model itself, but many more improvements, additions and clarifications to the supporting documentation.

6. The main changes to GSBPM between versions 4.0 and 5.0 are as follows:

- Phase 8 (Archive) has been removed, and incorporated into the over-arching process of data and metadata management, to reflect the view that archiving can happen at any stage in the statistical production process.
- A new sub-process: "Build or enhance dissemination components" has been added within the "Build" phase to reflect the growing importance of having a range of dissemination options.
- Several sub-processes have been re-named to improve clarity
- The descriptions of the sub-processes have been updated and expanded where necessary. The terminology used has been changed to be less survey-centric, in recognition of the growing use of non-survey sources (administrative data, big data etc.).

II. The Model

A. Understanding the Generic Statistical Business Process Model

7. GSBPM should be applied and interpreted flexibly. It is not a rigid framework in which all steps must be followed in a strict order, instead it identifies the possible steps in the statistical business process, and the inter-dependencies between them.

8. Although the presentation of GSBPM follows the logical sequence of steps in most statistical business processes, the elements of the model may occur in different orders in different circumstances. Also, some sub processes will be revisited a number of times forming iterative loops, particularly within the Process and Analyse phases.

9. GSBPM should therefore be seen more as a matrix, through which there are many possible paths. In this way GSBPM aims to be sufficiently generic to be widely applicable, and to encourage a standard view of the statistical business process, without becoming either too restrictive or too abstract and theoretical.

B. The structure

10. GSBPM comprises three levels:

- Level 0, the statistical business process;
- Level 1, the eight phases of the statistical business process;
- Level 2, the sub-processes within each phase.

11. A diagram showing the phases (level 1) and sub-processes (level 2) is included in Section IV. The sub-processes are described in detail in Section V.

12. GSBPM also recognises several over-arching processes that apply throughout the eight phases, and across statistical business processes. These can be grouped into two categories, those that have a statistical component, and those that are more general, and could apply to any sort of organisation. The first group ise considered to be more important in the context of this model, however the second group should also be recognised as they have (often indirect) impacts on several parts of the model.

13. Over-arching processes with a statistical component include the following. The first four are most closely related to the model. Quality and Metadata management are specifically highlighted in model diagrams and are elaborated further in Section VI.

- Quality management - This process includes quality assessment and control mechanisms. It recognises the importance of evaluation and feedback throughout the statistical business process;
- Metadata management - Metadata are generated and processed within each phase, there is, therefore, a strong requirement for a metadata management system to ensure the appropriate metadata retain their links with data throughout GSBPM. This includes process-independent considerations such as metadata custodianship and ownership, quality, archiving rules, preservation, retention and disposal;
- Data management - This includes process-independent considerations such as general data security, custodianship and ownership, data quality, archiving rules, preservation, retention and disposal;
- Process data management - This includes the management of data and metadata generated by and providing information on all parts of the statistical business process.
- Knowledge management - This ensures that statistical business processes are repeatable, mainly through the maintenance of process documentation;
- Statistical framework management - This includes developing standards, for example methodologies, concepts and classifications that apply across multiple processes;
- Statistical program management - This includes systematic monitoring and reviewing of emerging information requirements and emerging and changing data sources across all statistical domains. It may result in the definition of new statistical business processes or the redesign of existing ones;
- Provider management - This includes cross-process burden management, as well as topics such as profiling and management of contact information (and thus has particularly close links with statistical business processes that maintain registers);
- Customer management - This includes general marketing activities, promoting statistical literacy, and dealing with non-specific customer feedback.

14. More general over-arching processes include:

- Human resource management;
- Financial management;
- Project management;
- Legal framework management;
- Organisational framework management;
- Strategic planning.

C. Applicability

15. GSBPM is intended to apply to all activities undertaken by producers of official statistics, at both the national and international levels, which result in data outputs. It is designed to be independent of the data source, so it can be used for the description and quality assessment of processes based on surveys, censuses, administrative records, and other non-statistical or mixed sources.

16. Whilst the typical statistical business process includes the collection and processing of data to produce statistical outputs, GSBPM also applies to cases where existing data are revised or time-series are re-calculated, either as a result of improved source data, or a change in methodology. In these cases, the input data are the previously published statistics, which are then processed and analysed to produce revised outputs. In such cases, it is likely that several sub-processes and possibly some phases (particularly the early ones) would be omitted. Similarly, GSBPM can be applied to processes such as the compilation of National Accounts, and the typical processes of international statistical organisations

17. As well as being applicable for processes which result in statistics, GSBPM can also be applied to the development and maintenance of statistical registers, where the inputs are similar to those for statistical production (though typically with a greater focus on administrative data), and the outputs are typically frames or other data extractions, which are then used as inputs to other processes[6].

18. GSBPM should be seen as sufficiently flexible to apply in all of the above scenarios.

D. Using the Generic Statistical Business Process Model

19. GSBPM is a reference model. It is intended that GSBPM may be used by organisations to different degrees. An organisation may choose to either implement GSBPM directly or use it as the basis for developing an organisation specific adaption of the model. It may be used in some cases only as a model to which organisations refer when communicating internally or with other organisations to clarify discussion. The various scenarios for the use of GSBPM are all valid.

20. When organisations have developed organisation-specific adaptions of the GSBPM, they may make some specialisations to the model to fit their organisational context. The evidence so far suggests that these specialisations are not sufficiently generic to be included in the GSBPM itself.

21. In some cases it may be appropriate to group some of the elements of the model. For example, phases one to three could be considered to correspond to a single planning phase. In other cases, particularly for practical implementations, there may be a need to add one or more detailed levels to the structure presented below to separately identify different components of the sub-processes.

22. There may also be a requirement for a formal sign-off between phases, where the output from one phase is certified as suitable as input for the next. This sort of formal approval is implicit in the model, but may be implemented in many different ways depending on organisational requirements.

[6] See:
http://www1.unece.org/stat/platform/download/attachments/57835551/BR+meeting+paper2.doc

III. Relationships with other models and standards

23. The GSBPM was originally developed based on the Generic Business Process Model from Statistics New Zealand, supplemented by input from other statistical organisations with experience of statistical process modelling. However, a number of other related models and standards exist, both as a result of modernisation initiatives, and similar activities outside the scope of official statistics.

A. Common Statistical Production Architecture

24. Turning first to the links between GSBPM and the other frameworks and standards needed for statistical modernisation, it can be helpful to consider them in the context of enterprise architecture. Enterprise architecture relates business functions and processes to the information, applications and technology needed to run them. It is a way of describing what an organisation does, and how it does it, to try to identify how the organisation could improve quality and efficiency.

25. In the context of statistical modernisation, the aim is to align the enterprise architectures of different organisations, creating an "industry architecture" for the whole "official statistics industry". This approach is intended to facilitate collaboration, sharing and joint development of the components and services that are needed for the different parts of the statistical business process (defined in relation to the GSBPM). The result is the Common Statistical Production Architecture (CSPA), first released at the end of 2013[7].

B. Generic Statistical Information Model

26. The Generic Statistical Information Model (GSIM) is a reference framework for statistical information, designed to play an important part in modernising and streamlining official statistics at both national and international levels. It enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process. It provides a set of standardised, consistently described information objects, which are the inputs and outputs in the design and production of statistics. GSIM helps to explain significant relationships among the entities involved in statistical production, and can be used to guide the development and use of consistent implementation standards or specifications.

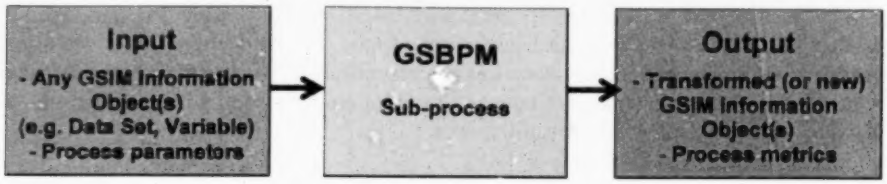
27. Like GSBPM, GSIM is one of the cornerstones for modernising official statistics and moving away from subject matter silos. GSIM is designed to allow for innovative approaches to statistical production to the greatest extent possible; for example, in the area of dissemination, where demands for agility and innovation are increasing. It also supports current approaches of producing statistics.

28. GSIM identifies around 110 information objects, examples include data sets, variables, statistical classifications, units, populations as well as the rules and parameters needed for production processes to run (for example, data editing rules).

29. GSIM and GSBPM are complementary models for the production and management of statistical information. As shown in the diagram below, GSIM helps describe GSBPM sub-processes by defining the information objects that flow between them, that are created

[7] See: http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home

in them, and that are used by them to produce official statistics. Inputs and outputs can be defined in terms of information objects, and are formalised in GSIM.



30. Greater value will be obtained from GSIM if it is applied in conjunction with GSBPM. Likewise, greater value will be obtained from GSBPM if it is applied in conjunction with GSIM. Nevertheless, it is possible (although not ideal) to apply one without the other. Similarly, both models support the implementation of CSPA, but can be applied regardless of whether that architectural framework is used or not.

31. In the same way that individual statistical business processes do not use all of the sub-processes described within GSBPM, it is very unlikely that all information objects in GSIM will be needed in any specific statistical business process.

32. Applying GSIM and GSBPM together can facilitate the building of efficient metadata driven systems, and help to harmonise statistical computing infrastructures.

C. Generic Longitudinal /Business Process Model

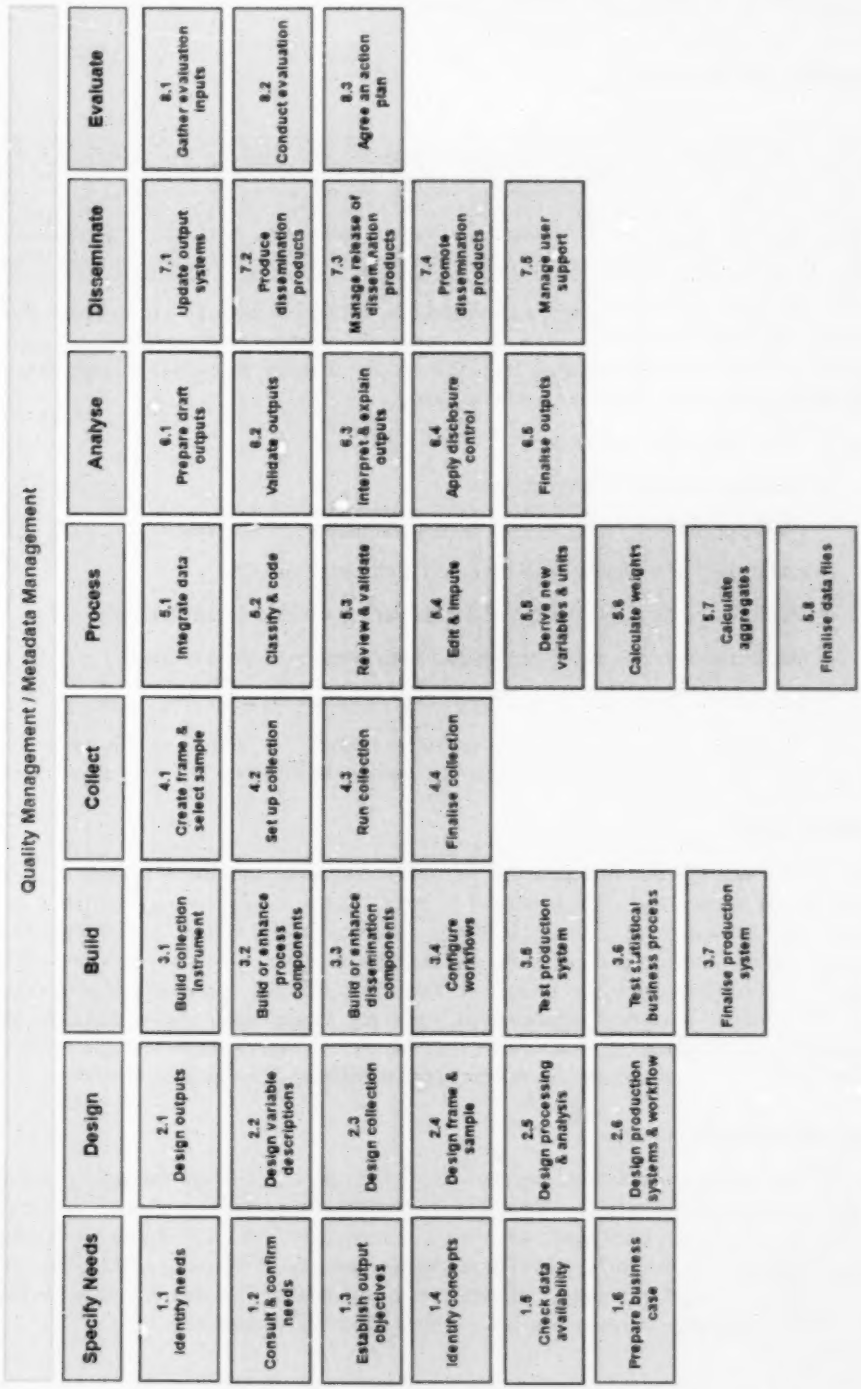
33. Looking outside the domain of official statistics, the social survey research community has developed the Generic Longitudinal Business Process Model (GLBPM)[8] "to provide a generic model that can serve as the basis for informing discussions across organisations conducting longitudinal data collections, and other data collections repeated across time".

34. Like GSBPM, GLBPM is a reference model against which actual business processes can be mapped. It is a reference model of the process of longitudinal and repeat cross-sectional data collection for research organisations. It describes the activities undertaken and maps these to their typical inputs and outputs.

35. GLBPM has many similarities to GSBPM, although it differs in some specific activities reflecting the different needs and practices of the statistical and social science research communities. GLBPM takes the approach of having a non-linear path through a matrix of alternatives directly from GSBPM.

[8] See: http://dx.doi.org/10.3886/DDILongitudinal05

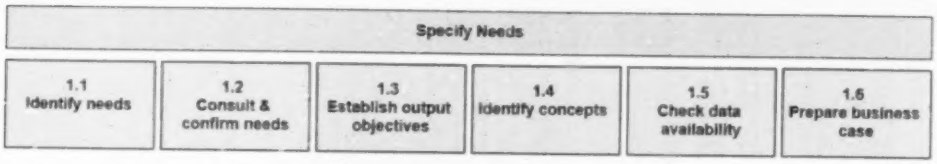
IV. Levels 1 and 2 of the Generic Statistical Business Process Model



V. Descriptions of phases and sub-processes

36. This section considers each phase in turn, identifying the various sub-processes within that phase, and describing their contents.

A. Specify Needs Phase



37. This phase is triggered when a need for new statistics is identified, or feedback about current statistics initiates a review. It includes all activities associated with engaging customers to identify their detailed statistical needs, proposing high level solution options and preparing business cases to meet these needs.

38. In this phase the organisation:

- identifies the need for the statistics;
- confirms, in more detail, the statistical needs of the stakeholders;
- establishes the high level objectives of the statistical outputs;
- identifies the relevant concepts and variables for which data are required;
- checks the extent to which current data sources can meet these needs;
- prepares the business case to get approval to produce the statistics.

39. This phase is broken down into six sub-processes. These are generally sequential, from left to right, but can also occur in parallel, and can be iterative. The sub-processes are:

1. Identify needs

40. This sub-process includes the initial investigation and identification of what statistics are needed and what is needed of the statistics. It may be triggered by a new information request, an environmental change such as a reduced budget. Action plans from evaluations of previous iterations of the process, or from other processes, might provide an input to this sub-process. It also includes consideration of practice amongst other (national and international) statistical organisations producing similar data, and in particular the methods used by those organisations. It may involve consideration of specific needs of different user communities, such as the disabled, or different ethnic groups.

2. Consult and confirm needs

41. This sub-process focuses on consulting with the stakeholders and confirming in detail the needs for the statistics. A good understanding of user needs is required so that the statistical organisation knows not only what it is expected to deliver, but also when, how, and, perhaps most importantly, why. For second and subsequent iterations of this phase, the main focus will be on determining whether previously identified needs have changed. This detailed understanding of user needs is the critical part of this sub-process.

3. Establish output objectives

42. This sub-process identifies the statistical outputs that are required to meet the user needs identified in sub-process 1.2 (Consult and confirm needs). It includes agreeing the suitability of the proposed outputs and their quality measures with users. Legal frameworks (e.g. relating to confidentiality), and available resources are likely to be constraints when establishing output objectives.

4. Identify concepts

43. This sub-process clarifies the required concepts to be measured by the business process from the point of view of the user. At this stage the concepts identified may not align with existing statistical standards. This alignment, and the choice or definition of the statistical concepts and variables to be used, takes place in sub-process 2.2.

5. Check data availability

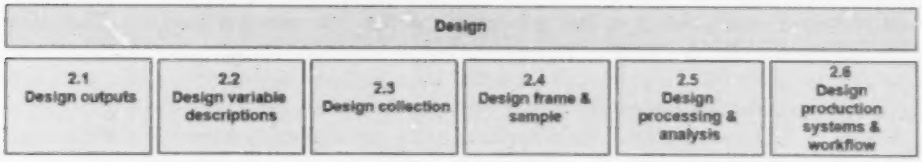
44. This sub-process checks whether current data sources could meet user requirements, and the conditions under which they would be available, including any restrictions on their use. An assessment of possible alternatives would normally include research into potential administrative or other non-statistical data sources, to determine whether they would be suitable for use for statistical purposes. When existing sources have been assessed, a strategy for filling any remaining gaps in the data requirement is prepared. This sub-process also includes a more general assessment of the legal framework in which data would be collected and used, and may therefore identify proposals for changes to existing legislation or the introduction of a new legal framework.

6. Prepare business case

45. This sub-process documents the findings of the other sub-processes in this phase in the form of a business case to get approval to implement the new or modified statistical business process. Such a business case would need to conform to the requirements of the approval body, but would typically include elements such as:

- A description of the "As-Is" business process (if it already exists), with information on how the current statistics are produced, highlighting any inefficiencies and issues to be addressed;
- The proposed "To-Be" solution, detailing how the statistical business process will be developed to produce the new or revised statistics;
- An assessment of costs and benefits, as well as any external constraints.

B. Design Phase



46. This phase describes the development and design activities, and any associated practical research work needed to define the statistical outputs, concepts, methodologies,

collection instruments[7] and operational processes. It includes all the design elements needed to define or refine the statistical products or services identified in the business case. This phase specifies all relevant metadata, ready for use later in the statistical business process, as well as quality assurance procedures. For statistical outputs produced on a regular basis, this phase usually occurs for the first iteration, and whenever improvement actions are identified in the Evaluate phase of a previous iteration.

47. Design activities make substantial use of international and national standards, in order to reduce the length and cost of the design process, and enhance to comparability and usability of outputs. Organisations are also encouraged to reuse or adapt design elements from existing processes. Additionally, outputs of design processes may form the basis for future standards at the organisation, national or international levels.

48. This phase is broken down into six sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

1. Design outputs

49. This sub-process contains the detailed design of the statistical outputs, products and services to be produced, including the related development work and preparation of the systems and tools used in the "Disseminate" phase. Disclosure control methods, as well as processes governing access to any confidential outputs are also designed here. Outputs should be designed to follow existing standards wherever possible, so inputs to this process may include metadata from similar or previous collections, international standards, and information about practices in other statistical organisations from sub-process 1.1 (Identify needs).

2. Design variable descriptions

50. This sub-process defines the statistical variables to be collected via the collection instrument, as well as any other variables that will be derived from them in sub-process 5.5 (Derive new variables and units), and any statistical classifications that will be used. It is expected that existing national and international standards will be followed wherever possible. This sub-process may need to run in parallel with sub-process 2.3 (Design collection), as the definition of the variables to be collected, and the choice of collection instrument may be inter-dependent to some degree. Preparation of metadata descriptions of collected and derived variables and classifications is a necessary precondition for subsequent phases.

3. Design collection

51. This sub-process determines the most appropriate collection method(s) and instrument(s). The actual activities in this sub-process will vary according to the type of collection instruments required, which can include computer assisted interviewing, paper questionnaires, administrative data interfaces and data integration techniques. This sub-process includes the design of collection instruments, questions and response templates (in conjunction with the variables and statistical classifications designed in sub-process 2.2 (Design variable descriptions)). It also includes the design of any formal agreements relating to data supply, such as memoranda of understanding, and confirmation of the legal basis for the data collection. This sub-process is enabled by tools such as question libraries (to facilitate the reuse of questions and related attributes), questionnaire tools (to enable the

[7] For GSBPM purposes, collection instruments are defined broadly to include any tool or routine to gather or extract data and metadata, from paper questionnaires to web-scraping tools. In GSIM version 1.1, collection instruments are "exchange channels" used for incoming information.

quick and easy compilation of questions into formats suitable for cognitive testing) and agreement templates (to help standardise terms and conditions). This sub-process also includes the design of process-specific provider management systems.

4. Design frame and sample

52. This sub-process only applies to processes which involve data collection based on sampling, such as through statistical surveys. It identifies and specifies the population of interest, defines a sampling frame (and, where necessary, the register from which it is derived), and determines the most appropriate sampling criteria and methodology (which could include complete enumeration). Common sources for a sampling frame are administrative and statistical registers, censuses and information from other sample surveys. This sub-process describes how these sources can be combined if needed. Analysis of whether the frame covers the target population should be performed. A sampling plan should be made: The actual sample is created in sub-process 4.1 (Create frame and select sample), using the methodology, specified in this sub-process.

5. Design processing and analysis

53. This sub-process designs the statistical processing methodology to be applied during the "Process" and "Analyse" phases. This can include specification of routines for coding, editing, imputing, estimating, integrating, validating and finalizing data sets.

6. Design production systems and workflow

54. This sub-process determines the workflow from data collection to dissemination, taking an overview of all the processes required within the whole statistical production process, and ensuring that they fit together efficiently with no gaps or redundancies. Various systems and databases are needed throughout the process. A general principle is to reuse processes and technology across many statistical business processes, so existing production solutions (e.g. services, systems and databases) should be examined first, to determine whether they are fit for purpose for this specific process, then, if any gaps are identified, new solutions should be designed. This sub-process also considers how staff will interact with systems, and who will be responsible for what and when.

C. Build Phase

Build						
3.1 Build collection instrument	3.2 Build or enhance process components	3.3 Build or enhance dissemination components	3.4 Configure workflows	3.5 Test production system	3.6 Test statistical business process	3.7 Finalise production system

55. This phase builds and tests the production solution to the point where it is ready for use in the "live" environment. The outputs of the "Design" phase direct the selection of reusable processes, instruments, information, and services that are assembled and configured in this phase to create the complete operational environment to run the process. New services are built by exception, created in response to gaps in the existing catalogue of services sourced from within the organisation and externally. These new services are constructed to be broadly reusable within the statistical production architecture.

56. For statistical outputs produced on a regular basis, this phase usually occurs for the first iteration, and following a review or a change in methodology or technology, rather than for every iteration.

57. It is broken down into seven sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

1. Build collection instrument

58. This sub-process describes the activities to build the collection instruments to be used during the "Collect" phase. The collection instrument is generated or built based on the design specifications created during the "Design" phase. A collection may use one or more modes to receive the data, e.g. personal or telephone interviews; paper, electronic or web questionnaires; SDMX hubs. Collection instruments may also be data extraction routines used to gather data from existing statistical or administrative data sets. This sub-process also includes preparing and testing the contents and functioning of that instrument (e.g. testing the questions in a questionnaire). It is recommended to consider the direct connection of collection instruments to the statistical metadata system, so that metadata can be more easily captured in the collection phase. Connection of metadata and data at the point of capture can save work in later phases. Capturing the metrics of data collection (paradata) is also an important consideration in this sub-process.

2. Build or enhance process components

59. This sub-process describes the activities to build new and enhance existing components and services needed for the "Process" and "Analyse" phases, as designed in the "Design" phase. Services may include dashboard functions and features, information services, transformation functions, workflow frameworks, provider and metadata management services.

3. Build or enhance dissemination components

60. This sub-process describes the activities to build new and enhance existing components and services needed for the dissemination of statistical products as designed in sub-process 2.1 (Design outputs). All types of dissemination components and services are included, from those that are used to produce traditional paper publications to those that provide web services, open data outputs, or access to micro-data.

4. Configure workflows

61. This sub-process configures the workflow, systems and transformations used within the statistical business processes, from data collection through to dissemination. It ensures that the workflow specified in sub-process 2.6 (Design production systems and workflow) works in practice.

5 Test production system

62. This sub-process is concerned with the testing of assembled and configured services and related workflows. It includes technical testing and sign-off of new programmes and routines, as well as confirmation that existing routines from other statistical business processes are suitable for use in this case. Whilst part of this activity concerning the testing of individual components and services could logically be linked with sub-process 3.2 (Build or enhance process components), this sub-process also includes testing of interactions between assembled and configured services, and ensuring that the production solution works as a coherent set processes, information and services.

6. Test statistical business process

63. This sub-process describes the activities to manage a field test or pilot of the statistical business process. Typically it includes a small-scale data collection, to test collection instruments, followed by processing and analysis of the collected data, to ensure the statistical business process performs as expected. Following the pilot, it may be

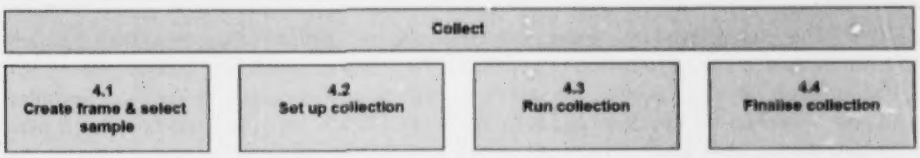
necessary to go back to a previous step and make adjustments to instruments, systems or components. For a major statistical business process, e.g. a population census, there may be several iterations until the process is working satisfactorily.

7. Finalise production systems

64. This sub-process includes the activities to put the assembled and configured processes and services, including modified and newly-created services into production ready for use by business areas. The activities include:

- producing documentation about the process components, including technical documentation and user manuals
- training the business users on how to operate the process
- moving the process components into the production environment, and ensuring they work as expected in that environment (this activity may also be part of sub-process 3.5 (Test production system)).

D. Collect Phase



65. This phase collects or gathers all necessary information (data and metadata), using different collection modes (including extractions from statistical, administrative and other non-statistical registers and databases), and loads them into the appropriate environment for further processing. Whilst it can include validation of data set formats, it does not include any transformations of the data themselves, as these are all done in the "Process" phase. For statistical outputs produced regularly, this phase occurs in each iteration.

66. The "Collect" phase is broken down into four sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

1. Create frame and select sample

67. This sub-process establishes the frame and selects the sample for this iteration of the collection, as specified in sub-process 2.4 (Design frame and sample). It also includes the coordination of samples between instances of the same statistical business process (for example to manage overlap or rotation), and between different processes using a common frame or register (for example to manage overlap or to spread response burden). Quality assurance and approval of the frame and the selected sample are also undertaken in this sub-process, though maintenance of underlying registers, from which frames for several statistical business processes are drawn, is treated as a separate business process. The sampling aspect of this sub-process is not usually relevant for processes based entirely on the use of pre-existing sources (e.g. administrative sources) as such processes generally create frames from the available data and then follow a census approach.

2. Set up collection

68. This sub-process ensures that the people, processes and technology are ready to collect data and metadata, in all modes as designed. It takes place over a period of time, as it includes the strategy, planning and training activities in preparation for the specific instance of the statistical business process. Where the process is repeated regularly, some

(or all) of these activities may not be explicitly required for each iteration. For one-off and new processes, these activities can be lengthy. This sub-process includes:

- preparing a collection strategy;
- training collection staff;
- ensuring collection resources are available e.g. laptops;
- agreeing terms with any intermediate collection bodies, e.g. sub-contractors for computer assisted telephone interviewing
- configuring collection systems to request and receive the data;
- ensuring the security of data to be collected;
- preparing collection instruments (e.g. printing questionnaires, pre-filling them with existing data, loading questionnaires and data onto interviewers' computers etc.).

69. For non-survey sources, this sub-process will include ensuring that the necessary processes, systems and confidentiality procedures are in place, to receive or extract the necessary information from the source.

3. Run collection

70. This sub-process is where the collection is implemented, with the different instruments being used to collect or gather the information, which may include raw micro-data or aggregates produced at the source, as well as any associated metadata. It includes the initial contact with providers and any subsequent follow-up or reminder actions. It may include manual data entry at the point of contact, or fieldwork management, depending on the source and collection mode. It records when and how providers were contacted, and whether they have responded. This sub-process also includes the management of the providers involved in the current collection, ensuring that the relationship between the statistical organisation and data providers remains positive, and recording and responding to comments, queries and complaints. For administrative and other non-statistical sources, this process is brief: the provider is either contacted to send the information, or sends it as scheduled. When the collection meets its targets, it is closed and a report on the collection is produced. Some basic validation of the structure and integrity of the information received may take place within this sub-process, e.g. checking that files are in the right format and contain the expected fields. All validation of the content takes place in the Process phase.

4. Finalise collection

71. This sub-process includes loading the collected data and metadata into a suitable electronic environment for further processing. It may include manual or automatic data take-on, for example using clerical staff or optical character recognition tools to extract information from paper questionnaires, or converting the formats of files received from other organisations. It may also include analysis of the process metadata (paradata) associated with collection to ensure the collection activities have met requirements. In cases where there is a physical collection instrument, such as a paper questionnaire, which is not needed for further processing, this sub-process manages the archiving of that material.

E. Process Phase

Process							
5.1 Integrate data	5.2 Classify & code	5.3 Review & validate	5.4 Edit & impute	5.5 Derive new variables & units	5.6 Calculate weights	5.7 Calculate aggregates	5.8 Finalise data files

72. This phase describes the cleaning of data and their preparation for analysis. It is made up of sub-processes that check, clean, and transform input data, so that they can be analysed and disseminated as statistical outputs. It may be repeated several times if necessary. For statistical outputs produced regularly, this phase occurs in each iteration. The sub-processes in this phase can apply to data from both statistical and non-statistical sources (with the possible exception of sub-process 5.6 (Calculate weights), which is usually specific to survey data).

73. The "Process" and "Analyse" phases can be iterative and parallel. Analysis can reveal a broader understanding of the data, which might make it apparent that additional processing is needed. Activities within the "Process" and "Analyse" phases may commence before the "Collect" phase is completed. This enables the compilation of provisional results where timeliness is an important concern for users, and increases the time available for analysis.

74. This phase is broken down into eight sub-processes, which may be sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

1. Integrate data

75. This sub-process integrates data from one or more sources. It is where the results of sub-processes in the "Collect" phase are combined. The input data can be from a mixture of external or internal data sources, and a variety of collection modes, including extracts of administrative data. The result is a set of linked data. Data integration can include:

- combining data from multiple sources, as part of the creation of integrated statistics such as national accounts
- matching / record linkage routines, with the aim of linking micro or macro data from different sources
- prioritising, when two or more sources contain data for the same variable, with potentially different values

76. Data integration may take place at any point in this phase, before or after any of the other sub-processes. There may also be several instances of data integration in any statistical business process. Following integration, depending on data protection requirements, data may be anonymised, that is stripped of identifiers such as name and address, to help to protect confidentiality.

2. Classify and code

77. This sub-process classifies and codes the input data. For example automatic (or clerical) coding routines may assign numeric codes to text responses according to a pre-determined classification scheme.

3. Review and validate

78. This sub-process examines data to try to identify potential problems, errors and discrepancies such as outliers, item non-response and miscoding. It can also be referred to

as input data validation. It may be run iteratively, validating data against predefined edit rules, usually in a set order. It may flag data for automatic or manual inspection or editing. Reviewing and validating can apply to data from any type of source, before and after integration. Whilst validation is treated as part of the "Process" phase, in practice, some elements of validation may occur alongside collection activities, particularly for modes such as web collection. Whilst this sub-process is concerned with detection of actual or potential errors, any correction activities that actually change the data are done in sub-process 5.4.

4. Edit and impute

79. Where data are considered incorrect, missing or unreliable, new values may be inserted in this sub-process. The terms editing and imputation cover a variety of methods to do this, often using a rule-based approach. Specific steps typically include:

- the determination of whether to add or change data;
- the selection of the method to be used;
- adding / changing data values;
- writing the new data values back to the data set, and flagging them as changed;
- the production of metadata on the editing and imputation process.

5. Derive new variables and units

80. This sub-process derives data for variables and units that are not explicitly provided in the collection, but are needed to deliver the required outputs. It derives new variables by applying arithmetic formulae to one or more of the variables that are already present in the dataset, or applying different model assumptions. This activity may need to be iterative, as some derived variables may themselves be based on other derived variables. It is therefore important to ensure that variables are derived in the correct order. New units may be derived by aggregating or splitting data for collection units, or by various other estimation methods. Examples include deriving households where the collection units are persons, or enterprises where the collection units are legal units.

6. Calculate weights

81. This sub process creates weights for unit data records according to the methodology created in sub-process 2.5 (Design processing and analysis). In the case of sample surveys, weights can be used to "gross-up" results to make them representative of the target population, or to adjust for non-response in total enumerations. In other situations, variables may need weighting for normalisation purposes.

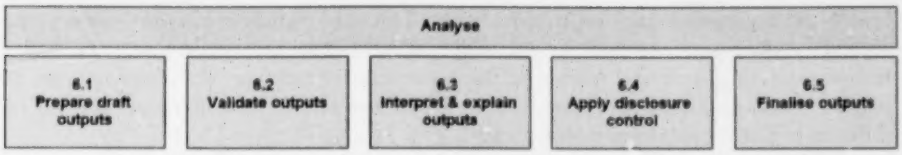
7. Calculate aggregates

82. This sub process creates aggregate data and population totals from micro-data or lower-level aggregates. It includes summing data for records sharing certain characteristics, determining measures of average and dispersion, and applying weights from sub-process 5.6 to derive appropriate totals. In the case of sample surveys, sampling errors may also be calculated in this sub-process, and associated to the relevant aggregates.

8. Finalise data files

83. This sub-process brings together the results of the other sub-processes in this phase and results in a data file (usually of macro-data), which is used as the input to the "Analyse" phase. Sometimes this may be an intermediate rather than a final file, particularly for business processes where there are strong time pressures, and a requirement to produce both preliminary and final estimates.

F. Analyse Phase



84. In this phase, statistical outputs are produced, examined in detail and made ready for dissemination. It includes preparing statistical content (including commentary, technical notes, etc.), and ensuring outputs are "fit for purpose" prior to dissemination to customers. This phase also includes the sub-processes and activities that enable statistical analysts to understand the statistics produced. For statistical outputs produced regularly, this phase occurs in every iteration. The "Analyse" phase and sub-processes are generic for all statistical outputs, regardless of how the data were sourced.

85. The "Analyse" phase is broken down into five sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. The sub-processes are:

1. Prepare draft outputs

86. This sub-process is where the data are transformed into statistical outputs. It includes the production of additional measurements such as indices, trends or seasonally adjusted series, as well as the recording of quality characteristics.

2. Validate outputs

87. This sub-process is where statisticians validate the quality of the outputs produced, in accordance with a general quality framework and with expectations. This sub-process also includes activities involved with the gathering of intelligence, with the cumulative effect of building up a body of knowledge about a specific statistical domain. This knowledge is then applied to the current collection, in the current environment, to identify any divergence from expectations and to allow informed analyses. Validation activi ies can include:

- checking that the population coverage and response rates are as required;
- comparing the statistics with previous cycles (if applicable);
- checking that the associated metadata and paradata (process metadata) are present and in line with expectations
- confronting the statistics against other relevant data (both internal and external);
- investigating inconsistencies in the statistics;
- performing macro editing;
- validating the statistics against expectations and domain intelligence.

3. Interpret and explain outputs

88. This sub-process is where the in-depth understanding of the outputs is gained by statisticians. They use that understanding to interpret and explain the statistics produced for this cycle by assessing how well the statistics reflect their initial expectations, viewing the statistics from all perspectives using different tools and media, and carrying out in-depth statistical analyses.

4. Apply disclosure control

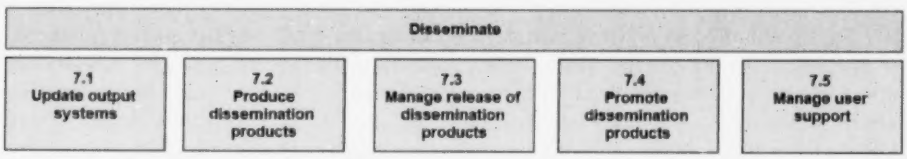
89. This sub-process ensures that the data (and metadata) to be disseminated do not breach the appropriate rules on confidentiality. This may include checks for primary and secondary disclosure, as well as the application of data suppression or perturbation techniques. The degree and method of disclosure control may vary for different types of outputs, for example the approach used for micro-data sets for research purposes will be different to that for published tables or maps.

5. Finalise outputs

90. This sub-process ensures the statistics and associated information are fit for purpose and reach the required quality level, and are thus ready for use. It includes:

- completing consistency checks;
- determining the level of release, and applying caveats;
- collating supporting information, including interpretation, commentary, technical notes, briefings, measures of uncertainty and any other necessary metadata;
- producing the supporting internal documents;
- pre-release discussion with appropriate internal subject matter experts;
- approving the statistical content for release.

G. Disseminate Phase



91. This phase manages the release of the statistical products to customers. It includes all activities associated with assembling and releasing a range of static and dynamic products via a range of channels. These activities support customers to access and use the outputs released by the statistical organisation.

92. For statistical outputs produced regularly, this phase occurs in each iteration. It is made up of five sub-processes, which are generally sequential, from left to right, but can also occur in parallel, and can be iterative. These sub-processes are:

1. Update output systems

93. This sub-process manages the update of systems where data and metadata are stored ready for dissemination purposes, including:

- formatting data and metadata ready to be put into output databases;
- loading data and metadata into output databases;
- ensuring data are linked to the relevant metadata.

94. Formatting, loading and linking of metadata should preferably mostly take place in earlier phases, but this sub-process includes a final check that all of the necessary metadata are in place ready for dissemination.

2. Produce dissemination products

95. This sub-process produces the products, as previously designed (in sub-process 2.1), to meet user needs. They could include printed publications, press releases and web sites. The products can take many forms including interactive graphics, tables, public-use micro-data sets and downloadable files. Typical steps include:

- preparing the product components (explanatory text, tables, charts, quality statements etc.);
- assembling the components into products;
- editing the products and checking that they meet publication standards.

3. Manage release of dissemination products

96. This sub-process ensures that all elements for the release are in place including managing the timing of the release. It includes briefings for specific groups such as the press or ministers, as well as the arrangements for any pre-release embargoes. It also includes the provision of products to subscribers, and managing access to confidential data by authorised user groups, such as researchers. Sometimes an organisation may need to retract a product, for example if an error is discovered. This is also included in this sub-process.

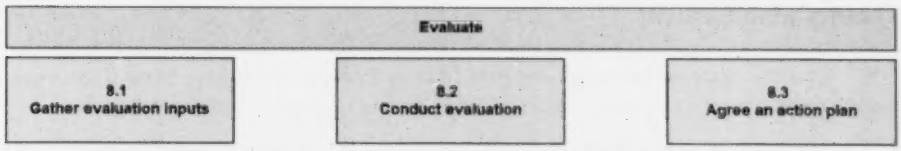
4. Promote dissemination products

97. Whilst marketing in general can be considered to be an over-arching process, this sub-process concerns the active promotion of the statistical products produced in a specific statistical business process, to help them reach the widest possible audience. It includes the use of customer relationship management tools, to better target potential users of the products, as well as the use of tools including web sites, wikis and blogs to facilitate the process of communicating statistical information to users.

5. Manage user support

98. This sub-process ensures that customer queries and requests for services such as micro-data access are recorded, and that responses are provided within agreed deadlines. These queries and requests should be regularly reviewed to provide an input to the over-arching quality management process, as they can indicate new or changing user needs.

H. Evaluate Phase



99. This phase manages the evaluation of a specific instance of a statistical business process, as opposed to the more general over-arching process of statistical quality management described in Section VI. It logically takes place at the end of the instance of the process, but relies on inputs gathered throughout the different phases. It includes evaluating the success of a specific instance of the statistical business process, drawing on a range of quantitative and qualitative inputs, and identifying and prioritising potential improvements.

100. For statistical outputs produced regularly, evaluation should, at least in theory occur for each iteration, determining whether future iterations should take place, and if so,

whether any improvements should be implemented. However, in some cases, particularly for regular and well established statistical business processes, evaluation may not be formally carried out for each iteration. In such cases, this phase can be seen as providing the decision as to whether the next iteration should start from the Specify Needs phase, or from some later phase (often the Collect phase).

101. This phase is made up of three sub-processes, which are generally sequential, from left to right, but which can overlap to some extent in practice. These sub-processes are:

1. Gather evaluation inputs

102. Evaluation material can be produced in any other phase or sub-process. It may take many forms, including feedback from users, process metadata (paradata), system metrics, and staff suggestions. Reports of progress against an action plan agreed during a previous iteration may also form an input to evaluations of subsequent iterations. This sub-process gathers all of these inputs, and makes them available for the person or team producing the evaluation.

2. Conduct evaluation

103. This sub-process analyses the evaluation inputs and synthesises them into an evaluation report. The resulting report should note any quality issues specific to this iteration of the statistical business process, and should make recommendations for changes if appropriate. These recommendations can cover changes to any phase or sub-process for future iterations of the process, or can suggest that the process is not repeated.

3. Agree an action plan

104. This sub-process brings together the necessary decision-making power to form and agree an action plan based on the evaluation report. It should also include consideration of a mechanism for monitoring the impact of those actions, which may, in turn, provide an input to evaluations of future iterations of the process.

VI. Over-arching processes

105. GSBPM also recognises several over-arching processes that apply throughout the production phases, and across statistical business processes. Some of these over-arching processes are listed in paragraph 13 and 14. The processes of quality management and metadata management are further elaborated in this Section.

A. Quality management

106. Quality concerns organisations, processes and products. In the present framework, quality management over-arching process refers mainly to product and process quality.

107. The main goal of quality management within the statistical business process is to understand and manage the quality of the statistical products. There is general agreement among statistical organisations that quality should be defined according to the ISO 9000-2005 standard: "The degree to which a set of inherent characteristics fulfils requirements"[10]. Thus, product quality is a complex and multi-faceted concept, usually defined in terms of several quality dimensions. The dimensions of quality that are considered most important

[10] ISO 9000:2005, Quality management systems -- Fundamentals and vocabulary. International Organization for Standardization

depend on user perspectives, needs and priorities, which vary between processes and across groups of users.

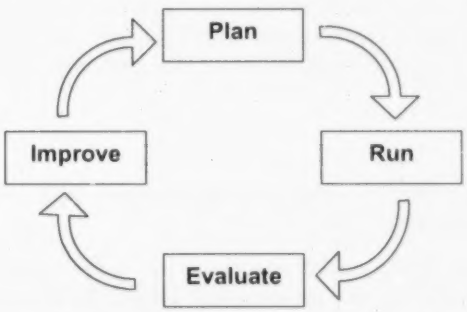
108. In order to improve the product quality, quality management should be present throughout the statistical business process model. It is closely linked to Phase 8 (Evaluate), which has the specific role of post-evaluating individual instances of a statistical business process. However, quality management has both a deeper and broader scope. As well as evaluating iterations of a process, it is also necessary to evaluate separate phases and sub-processes, ideally each time they are applied, but at least according to an agreed schedule. Metadata generated by the different sub-processes themselves are also of interest as an input for process quality management. These evaluations can apply within a specific process, or across several processes that use common components.

109. In addition, a fundamental role in quality management is played by the set of quality control actions that should be implemented within the sub-processes to prevent and monitor errors. The strategy could be reported in a quality assurance plan.

110. Within an organisation, quality management will usually refer to a specific quality framework, and may therefore take different forms and deliver different results within different organisations. The current multiplicity of quality frameworks enhances the importance of the benchmarking and peer review approaches to evaluation, and whilst these approaches are unlikely to be feasible for every iteration of every part of every statistical business process, they should be used in a systematic way according to a pre-determined schedule that allows for the review of all main parts of the process within a specified time period[11].

111. Broadening the field of application of the quality management over-arching process, evaluation of groups of statistical business processes can also be considered, in order to identify potential duplication or gaps.

112. All evaluations result in feedback, which should be used to improve the relevant process, phase or sub-process, creating a quality loop.



113. Examples of quality management activities include:

- Setting and maintaining of the quality framework;
- Setting of global quality criteria;
- Setting process quality targets and monitoring compliance;
- Seeking and analysing user feedback;

[11] A suitable global framework is the National Quality Assurance Framework developed by a global expert group under the United Nations Statistical Commission. See: http://unstats.un.org/unsd/dnss/QualityNQAF/nqaf.aspx

- Reviewing operation and documenting lessons learned;
- Examining process metadata and quality indicators;
- Internal or external auditing on statistical processes.

114. Quality management also involves institutional and organisational factors. Such factors are included in other GSBPM over-arching processes (e.g. Human resources management, Statistical programme management) although they can have an impact on quality.

B. Metadata management

115. Good metadata management is essential for the efficient operation of statistical business processes. Metadata are present in every phase, either created or carried forward from a previous phase. In the context of this model, the emphasis of the over-arching process of metadata management is on the creation, use and archiving of statistical metadata, though metadata on the different sub-processes themselves are also of interest, including as an input for quality management. The key challenge is to ensure that these metadata are captured as early as possible, and stored and transferred from phase to phase alongside the data they refer to. Metadata management strategy and systems are therefore vital to the operation of this model, and these can be facilitated by GSIM.

116. GSIM is a reference framework of information objects, which enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process. GSIM supports a consistent approach to metadata, facilitating the primary role for metadata envisaged in Part A of the Common Metadata Framework[12] "Statistical Metadata in a Corporate Context", that is, that metadata should uniquely and formally define the content and links between objects and processes in the statistical information system.

117. Part A of the Common Metadata Framework also identifies the following sixteen core principles for metadata management, all of which are intended to be covered in the over-arching Metadata Management process, and taken into the consideration when designing and implementing a statistical metadata system. The principles are presented in four groups:

1. Metadata handling

(a) *Statistical Business Process Model*: Manage metadata with a focus on the overall statistical business process model.

(b) *Active not passive*: Make metadata active to the greatest extent possible. Active metadata are metadata that drive other processes and actions. Treating metadata this way will ensure they are accurate and up-to-date.

(c) *Reuse*: Reuse metadata where possible for statistical integration as well as efficiency reasons

(d) *Versions*: Preserve history (old versions) of metadata.

2. Metadata Authority

(a) *Registration*: Ensure the registration process (workflow) associated with each metadata element is well documented so there is clear identification of ownership, approval status, date of operation, etc.

[12] See: http://www.unece.org/stats/cmf/PartA.html

(b) *Single source*: Ensure that a single, authoritative source ('registration authority') for each metadata element exists.

(c) *One entry/update*: Minimise errors by entering once and updating in one place.

(d) *Standards variations*: Ensure that variations from standards are tightly managed/approved, documented and visible.

3. Relationship to statistical cycle/process

(a) *Integrity*: Make metadata-related work an integral part of business processes across the organisation.

(b) *Matching metadata*: Ensure that metadata presented to the end-users match the metadata that drove the business process or were created during the process.

(c) *Describe flow*: Describe metadata flow with the statistical and business processes (alongside the data flow and business logic).

(d) *Capture at source*: Capture metadata at their source, preferably automatically as a by-product of other processes.

(e) *Exchange and use*: Exchange metadata and use them for informing both computer based processes and human interpretation. The infrastructure for exchange of data and associated metadata should be based on loosely coupled components, with a choice of standard exchange languages, such as XML.

VII. Other uses of the Generic Statistical Business Process Model

118. The original aim of GSBPM was to provide a basis for statistical organisations to agree on standard terminology to aid their discussions on developing statistical metadata systems and processes. However, as the model has developed, it has become increasingly apparent that it can be used for many other purposes, in particular related to modernisation of official statistics. A number of papers describing actual and potential uses of the GSBPM are available on the UNECE wiki platform[13]. The list below aims to highlight some current uses, and to inspire further ideas on how GSBPM can be used in practice.

- Harmonizing statistical production architectures - GSBPM can be seen as a model for an operational view of statistical computing architecture. It identifies the key components of the statistical business process, promotes standard terminology and standard ways of working across statistical business processes. It is a key enabler of the Common Statistical Production Architecture[14].
- Facilitating the sharing of statistical software - Linked to the point above, the GSBPM defines the components of statistical processes in a way that not only encourages the sharing of software tools between statistical business processes, but also facilitates sharing between different statistical organisations that apply the model. It has been used to "classify" software available for sharing in the inventory compiled by the Sharing Advisory Board[15].

[13] See: http://www1.unece.org/stat/platform/display/metis/Papers+about+the+GSBPM

[14] See: http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home

[15] See: http://www1.unece.org/stat/platform/display/msis/Software+Inventory

- Describing which standards are or could be used for different phases of the statistical production process. For example, Annex 2 of SDMX 2.1 User Guide[16] explores how SDMX applies to statistical work in the context of a business process model.
- Providing a framework for process quality assessment and improvement - If a benchmarking approach to process quality assessment is to be successful, it is necessary to standardise processes as much as possible. GSBPM provides a mechanism to facilitate this.
- Better integrating work on statistical metadata and quality - Linked to the previous point, the common framework provided by GSBPM can help to integrate international work on statistical metadata with that on data quality by providing a common framework and common terminology to describe the statistical business process.
- Providing the underlying model for methodological standards frameworks - Methodological standards can be linked to the phase(s) or sub-process(es) they relate to and can then be classified and stored in a structure based on GSBPM.
- Providing a structure for documentation of statistical processes - GSBPM can provide a structure for organizing and storing documentation within an organisation, promoting standardisation and the identification of good practices.
- Providing a framework for building organisational capability - GSBPM can be used to develop a framework assess the knowledge and capability that already exists within an organisation, and to identify the gaps that need to be filled to improve operational efficiency.
- Providing an input to high-level corporate work planning.
- Developing a business process model repository - Statistics New Zealand has developed a database to store process modelling outputs and allow them to be linked to their statistical business process model. They also plan to develop a Business Process Modelling Community of Practice - i.e. a regular forum to build knowledge of process modelling, to promote their business process model and increase understanding of it, and to discuss process modelling and models as enablers for process improvement.
- Measuring operational costs - GSBPM can be used as a basis for measuring the costs of different parts of the statistical business process. This helps to target modernisation activities to improve the efficiency of the parts of the process that are most costly.
- Measuring system performance - Related to the point above on costs, GSBPM can also be used to identify components that are not performing efficiently, that are duplicating each other unnecessarily, or that require replacing. Similarly it can identify gaps for which new components should be developed.

[16] See: http://sdmx.org/index.php?page_id=38

Annex 1– List of acronyms

Note: this list covers selected key abbreviations used in this paper. For a more comprehensive glossary of terms related to the statistical production process, see GSIM documentation.

CMF - Common Metadata Framework: A set of resources relating to the use of metadata by statistical organisations, including information on standards and best practices. See http://www.unece.org/stats/cmf/.

CSPA - Common Statistical Production Architecture: An industry architecture which brings together the GSBPM and GSIM, in addition to new frameworks about Statistical Services to create an agreed top level description of the 'system' of producing statistics which is in alignment with the modernisation initiative.

DDI - Data Documentation Initiative: An international standard for describing data from the social, behavioural, and economic sciences.

GLBPM - Generic Longitudinal /Business Process Model: A model based on the GSBPM, developed by the social survey research community.

GSBPM - Generic Statistical Business Process Model: A flexible tool to describe and define the set of business processes needed to produce official statistics.

GSIM - Generic Statistical Information Model: A reference framework of information objects, which enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process.

HLG - The High-Level Group for the Modernisation of Statistical Production and Services

METIS - The "brand name" for work on Statistical Metadata under the Conference of European Statisticians.

OECD - Organisation for Economic Cooperation and Development

SDMX - Statistical Data and Metadata eXchange: A set of technical standards and content-oriented guidelines, together with an IT architecture and tools, to be used for the efficient exchange and sharing of statistical data and metadata.

UNECE - United Nations Economic Commission for Europe

XML - eXtensible Mark-up Language: A language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable

United Nations ECE/CES/2014/2

Economic and Social Council

Distr.: General
28 January 2014

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-second plenary session
Paris, 9-11 April 2014
Item 5 of the provisional agenda
Work of the High-level Group for the Modernization of Statistical Production and Services

Generic Statistical Information Model (GSIM): Communication paper for a general statistical audience

Prepared by the High-Level Group for the Modernization of Statistical Production and Services

Summary

This paper introduces the Generic Statistical Information Model to people working in statistical organisations. It outlines the benefits of the model as well as how the adoption of the model might impact staff in statistical organisations. The paper also discusses the interaction of the Generic Statistical Information Model and other frameworks and standards such as the Generic Statistical Business Process Model, the Data Documentation Initiative and the Statistical Data and Metadata Exchange standard. More information on this information model is available in the Generic Statistical Information Model Specification document, excerpts provided in the Annex, and on the related wiki page[1] that might be useful to organisations adopting this information model as a corporate standard.

The document is submitted to the Conference of European Statisticians for information.

[1] http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

GE.14-20416

1420416

Please recycle

Contents

		Page
I.	Introduction	3
II.	Scope	4
III.	What is Generic Statistical Information Model?	4
IV.	Benefits of the Generic Statistical Information Model for the organization as a whole	6
	A. Long term benefits	7
	B. Immediate benefits	7
V.	Generic Statistical Information Model and Generic Statistical Business Process Model	8
VI.	What does it mean for me?	9
	A. The business view	9
	B. The information technology view	11
VII.	Statistical Data and Metadata Exchange standard, Data Documentation Initiative and other standards	12
Annex - Excerpts from Generic Statistical Information Model Specification document		14
	A. Information in the Statistical Business Process	14
	B. Foundational Information	25

I. Introduction

1. Across the world statistical organizations undertake similar activities albeit with variations in the processes each uses. Each of these activities use and produce similar information (for example all organizations use classifications, create data sets and disseminate information). Although the information used by statistical organizations is at its core the same, all organizations tend to describe this information slightly differently (and often in different ways within each organization). In the past, there was no common way to describe the information that is used. This makes it difficult to communicate clearly within and between statistical organizations and without this there was no foundation for in-depth collaboration, standardization, or the sharing of tools and methods.

2. The Generic Statistical Information Model (GSIM) is the first internationally endorsed reference framework for statistical information. This overarching conceptual framework will play an important part in modernizing, streamlining and aligning the standards and production associated with official statistics at both national and international levels.

3. GSIM is a reference framework of information objects, which enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process. It provides a set of standardized, consistently described information objects, which are the inputs and outputs in the design and production of statistics. As a reference framework, GSIM helps to explain significant relationships among the entities involved in statistical production, and can be used to guide the development and use of consistent implementation standards or specifications.

4. GSIM is one of the cornerstones for modernizing official statistics and moving away from subject matter silos. It is a key element of the strategic vision prepared by the High-Level Group for the Modernization of Statistical Production and Services (HLG), and endorsed by the Conference of European Statisticians[2].

5. The modernization of statistical production is needed in order for statistical organizations to remain relevant and flexible in a dynamic and competitive information environment. It is hoped that statistical organizations will adopt and implement GSIM and the common language it provides. However, a model alone cannot transform an organization or its processes. In order to meet the future needs of statistical organizations, GSIM is designed to allow for innovative approaches to statistical production to the greatest extent possible. It is one of the main foundations of the Common Statistical Production Architecture[3], a collaborative initiative to design common and interchangeable services with standard interfaces to support standardisation and modernisation. At the same time, GSIM supports current ways of producing statistics.

6. This paper provides an introduction to GSIM, summarizing the key points for a relatively general statistical audience. For more technical detail, please see the specification document and related material, available on the UNECE web site[4].

[2] See: www1.unece.org/stat/platform/display/hlgbas

[3] See: http://www1.unece.org/stat/platform/display/CSPA

[4] See:
http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model+(GSIM)

II. Scope

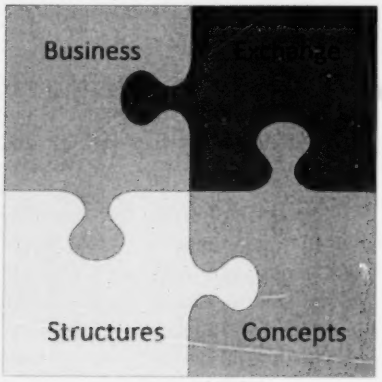
7. GSIM provides the information object framework supporting all statistical production processes such as those described in the Generic Statistical Business Process Model (GSBPM)[5], giving the information objects agreed names, defining them, specifying their essential properties, and indicating their relationships with other information objects. It does not, however, make assumptions about the standards or technologies used to implement the model.

8. GSIM does not include information objects related to business functions within an organization such as human resources, finance, or legal functions, except to the extent that this information is used directly in statistical production.

III. What is Generic Statistical Information Model?

9. GSIM contains objects which specify information about the real world – 'information objects'. Examples include data and metadata (such as classifications) as well as the rules and parameters needed for production processes to run (for example, data editing rules). GSIM identifies around 110 information objects, which are grouped into four top-level groups, and are explained in more detail in the specification documentation.

Figure 1.
GSIM Top-level information object Groups



10. The four top-level groups are described below:

(a) The *business group* is used to capture the designs and plans of statistical programmes, and the processes undertaken to deliver those programmes. This includes the identification of a Statistical Need, the Business Processes that comprise the Statistical Programme and the evaluations of them.

(b) The *exchange group* is used to catalogue the information that comes in and out of a statistical organization via Exchange Channels. It includes objects that describe the collection and dissemination of information.

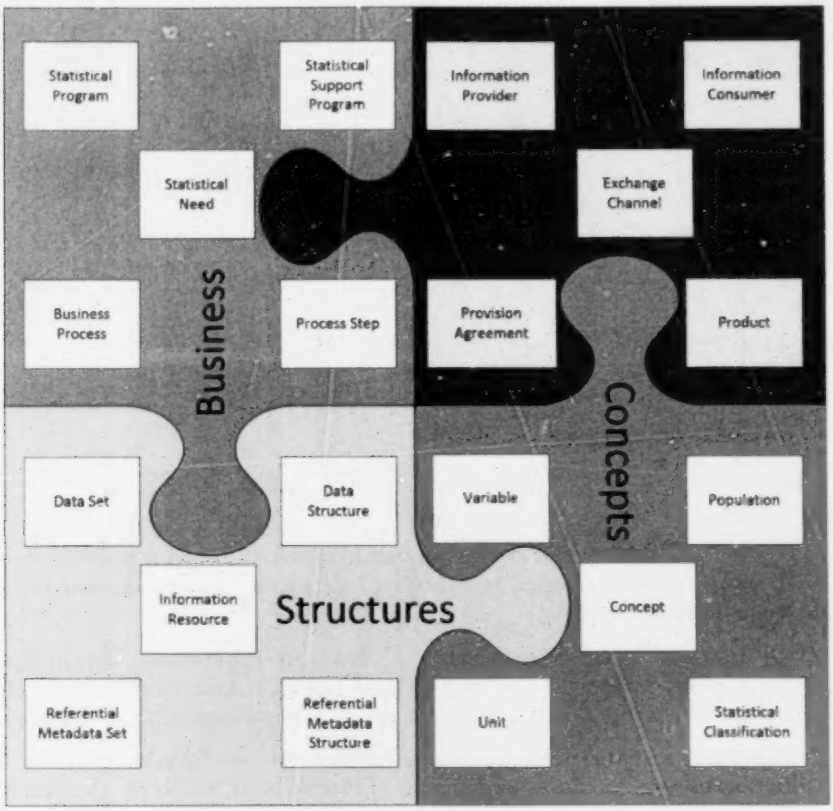
(c) The *concepts group* is used to define the meaning of data, providing an understanding of what the data are measuring.

(d) The *structures group* is used to describe and define the terms used in relation to information and its structure.

[5] See: www.unece.org/stats/gsbpm, document number (ECE/CES/2014/1).

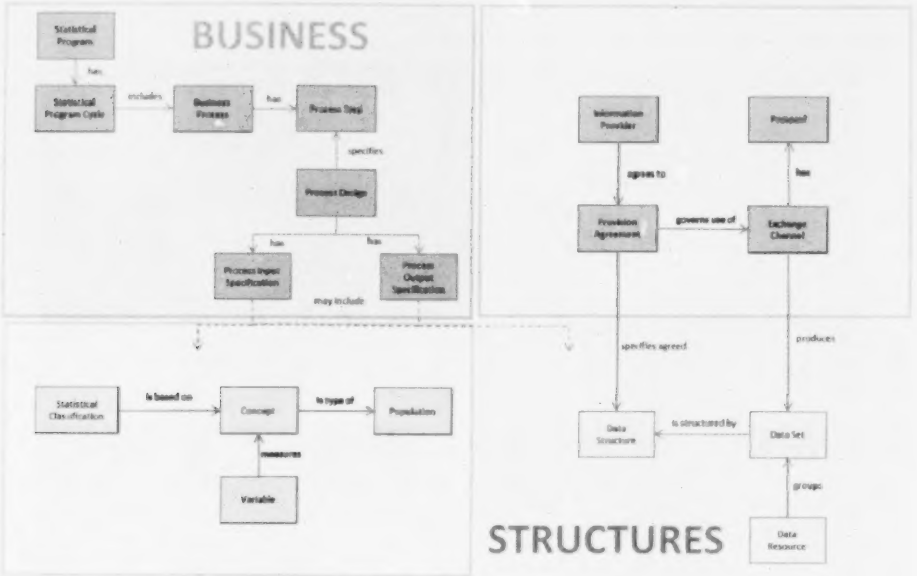
11. Figure 2 shows a simplified view of the information objects identified in GSIM. It gives users examples of the objects that are in each of the four top-level groups.

Figure 2.
Simplified view of GSIM information objects



12. Figure 3 shows another view of one part of GSIM. This is a slightly more technical view, but still intended to be accessible by a relatively wide audience. Both figures 2 and 3 can be used as a means for communication with users who are interested in examples of the objects and relationships in GSIM.

Figure 3.
Alternate simplified view of GSIM information objects



13. Figure 3 gives an example of GSIM information objects that tell a story about some of the information that is important in a statistical organisation. Information objects in the GSIM model are given in italics.

> "A statistical organization initiates a *Statistical Programme*. The *Statistical Programme* corresponds to an on-going activity such as a survey or an output series and has a *Statistical Program Cycle* (for example it repeats quarterly or annually).
>
> The Statistical Programme Cycle will include a set of Business Processes. The Business Processes consist of a number of Process Steps which are specified by a Process Design. These Process Designs have Process Input Specifications and Process Output Specifications. The specifications will often be pieces of information that refer to Concepts and Structures (for example, Statistical Classification, Variable, Population, Data Structure, and Data Set).
>
> If, for example, the *Business Process* is related to the collection of data, there will be an *Information Provider* who agrees to provide the statistical organisation with data (via a *Provision Agreement*). This *Provision Agreement* specifies an agreed *Data Structure* and governs the *Exchange Channel* used for the incoming information. The *Exchange Channel* could be a *Questionnaire* or an *Administrative Register*. It will receive the information via a particular mechanism (*Protocol*) such as an interview or a data file exchange.
>
> The *Data Set* produced by the *Exchange Channel* will be stored in a *Data Resource* and structured by a *Data Structure*."

IV. Benefits of the Generic Statistical Information Model for the organization as a whole

14. It is intended that GSIM may be used by organizations to different degrees. It may be used in some cases only as a model to which organizations refer when communicating internally or with other organizations to clarify discussion. In other cases an organization

may choose to implement GSIM as the information model that defines their operating environment. Various scenarios for the use of GSIM are valid, although those organizations that make use of GSIM to its fullest extent may expect to realize the greatest benefits.

A. Long term benefits

15. GSIM provides a set of standardized information objects, which are the inputs and outputs in the design and production of statistics. By defining objects common to all statistical production, regardless of subject matter, GSIM enables statistical organizations to rethink how their business could be more efficiently organized.

16. GSIM could be used to direct future investment towards areas of statistical production where the common need is greatest. It could also enable some degree of specialization within the international statistical community. For example, some organizations could specialize in seasonal adjustment, time series analysis or data validation, and other organizations could take advantage of this expertise.

17. Implementation of GSIM, in combination with GSBPM, will lead to more important advantages. GSIM could:

- Create an environment prepared for reuse and sharing of methods, components and processes;
- Provide the opportunity to implement rule based process control, thus minimizing human intervention in the production process;
- Facilitate generation of economies of scale through development of common tools by the community of statistical organizations.

B. Immediate benefits

18. A significant benefit of using GSIM is that it provides a common language to improve communication at different levels:

- Between the different roles in statistical production (business and information technology experts);
- Between the different statistical subject matter domains;
- Between statistical organizations at national and international levels.

19. Improving communication will result in a more efficient exchange of data and metadata within and between statistical organizations, and also with external users and suppliers.

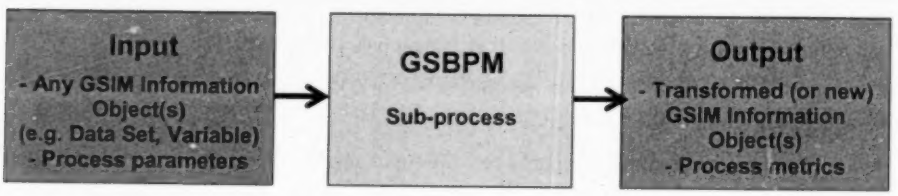
20. GSIM can be used by organizations now to:

- Build capability among staff by using GSIM as a teaching aid that provides a simple easy to understand view of complex information and clear definitions;
- Validate existing information systems and compare with emerging international best practice and where appropriate leverage off international expertise;
- Guide development or updating of international or local standards to ensure they meet the broadest needs of the international statistical community.

V. Generic Statistical Information Model and Generic Statistical Business Process Model

21. GSIM and GSBPM are complementary models for the production and management of statistical information. GSBPM models the statistical production process and identifies the activities undertaken by producers of official statistics that result in information outputs. These activities are broken down into sub-processes, such as "Impute" and "Calculate aggregates". As shown in Figure 6, GSIM helps describe GSBPM sub-processes by defining the information objects that flow between them, that are created in them, and that are used by them to produce official statistics.

Figure 4.
The relation of GSIM and GSBPM



22. Greater value will be obtained from GSIM if it is applied in conjunction with GSBPM. Likewise, greater value will be obtained from GSBPM if it is applied in conjunction with GSIM. Nevertheless, it is possible (although not ideal) to apply one without the other. In the same way that individual statistical business processes do not use all of the sub-processes described within GSBPM, it is very unlikely that all information objects in the GSIM will be needed in any specific statistical business process.

23. Good metadata management is essential for the efficient operation of statistical business processes. Metadata are present in every phase of GSBPM, either created, updated or carried forward unchanged from a previous phase. In the context of GSBPM, the emphasis of the over-arching process of metadata management is on the creation, updating, use and reuse of metadata. Metadata management strategies and systems are therefore vital to the operation of GSBPM, and are facilitated by GSIM.

24. Applying GSIM together with GSBPM (or an organization-specific equivalent) can:

- Facilitate the building of efficient metadata driven collection, processing, and dissemination systems;
- Help harmonize statistical computing infrastructures.

25. GSIM supports a consistent approach to metadata, facilitating the primary role for metadata envisaged in Part A of the Common Metadata Framework "Statistical Metadata in a Corporate Context"[6], that is, that metadata should *uniquely and formally define the content and links between objects and processes in the statistical information system.*

[6] http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework

VI. What does it mean for me?

A. The business view

26. GSIM will help you to improve your communication with colleagues (both locally and internationally).

27. Communication of subject matter between domains is often poor, making the sharing of concepts, variables, and design components difficult without a complex mapping exercise. GSIM can serve as a common language and will ease communication between:

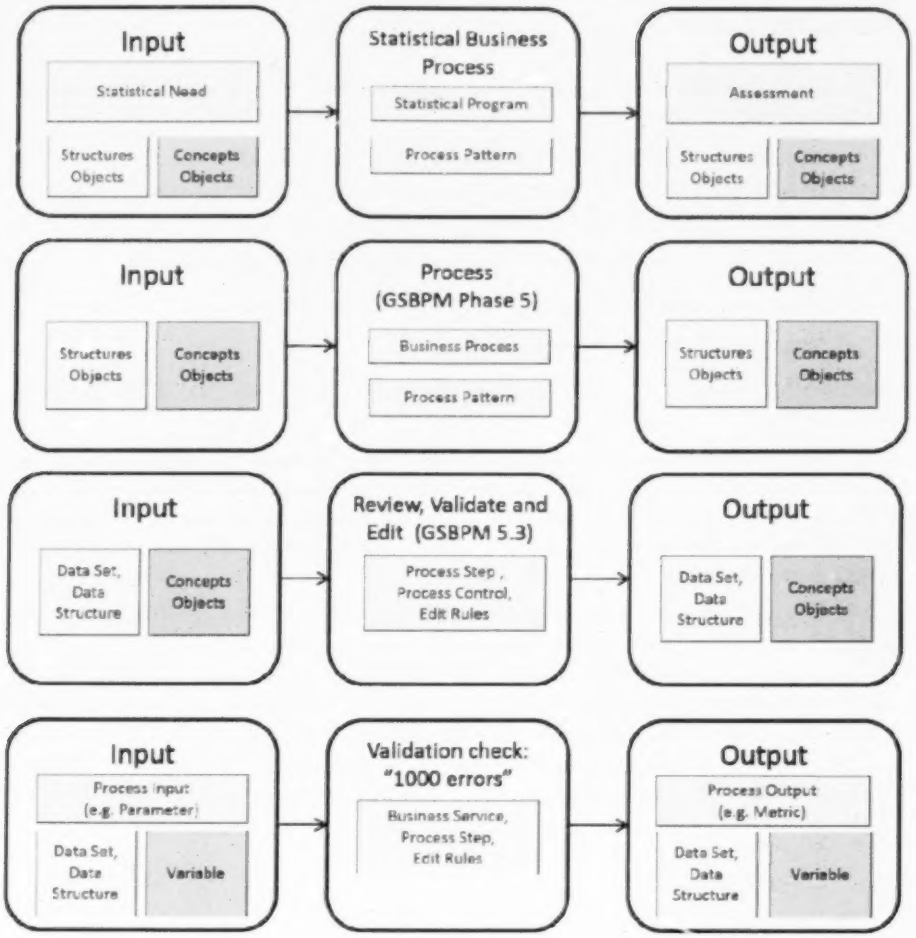
- Subject matter specialists, methodologists and information technologists;
- Statisticians in different domains of a statistical organization;
- Statisticians in different organizations.

28. GSIM will help you design and understand your processes (and their inputs and outputs) better.

29. For a production cycle, a statistician can design the input and the output, and the process in-between. In GSIM terms, the output and the input can be designed in terms of structures and concepts information objects, and the process in-between can be designed using the business information objects. The structures and concepts objects are provided by subject matter specialists.

30. As seen in Figure 4, if the GSBPM is considered as a frame of reference for statistical production processes, the first level can be considered as equivalent to the statistical production process as a whole. The next level corresponds to a phase of the statistical production process (for example the "Process" phase of the GSBPM). The third level corresponds to a sub-process (for example sub-process 5.3 of the GSBPM – Review and validate). The fourth level consists of the individual building blocks within the sub-process, such as detecting financial values that might be expressed in thousands rather than units.

Figure 5.
GSIM information objects in context of GSBPM



31. An important issue for statisticians is the problem of single-use design components, which are often recreated or at least modified for each production cycle. GSIM facilitates the description of inputs and outputs at each level of the GSBPM, following the same pattern thus providing a consistent structure to design statistical processes. It supports the design, specification and implementation of harmonized methods and standard technology to create a generalized statistical production system.

32. Using GSIM will enable producing reusable and flexible process building blocks which can be used by statisticians to produce final products of varying complexity, facilitating the production of a wider variety of products and responding more easily to changing client needs.

33. The use of GSIM will reduce workloads as many processes can be repurposed and reused. This means less time spent on repetitive work and more time for innovation.

34. In the long term, GSIM will make statisticians less reliant on information technologists.

35. Statisticians are very much concerned today about the applicability, usability and stability of their methods and technical solutions. In the "stove-pipe" approach to statistical

production, subject matter is heavily dependent upon the information technologists in the design, build and production of statistical systems.

36. Statisticians will gain greater control over the design of their processes making them more self-supporting in the design and production of their statistics.

37. Production will be based upon more standardized applications that are more robust to change and less vulnerable to changing personnel. An increase in the use of standardized applications, which can easily be shared across domains, will enable statisticians to more easily work in different domains.

B. The information technology view

38. A main concern for information technologists is the duplication of effort due to the "stove-pipe" organization of statistical production. Unstable and differing requirements from these "stove-pipes" lead to tailor made one–off solutions, whilst a high turnover of IT staff can result in poorly documented and non-standard applications.

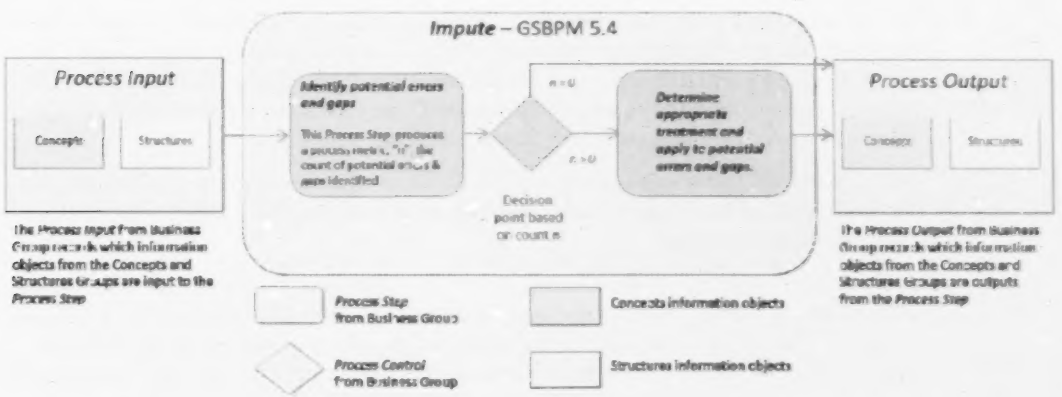
39. The introduction of GSIM both at the national and at the international level can already bring short term benefits for information technology specialists. GSIM will provide a common language for information technologists to talk to clients and colleagues both locally and internationally.

40. At the national level, statisticians will become more self-supporting in the design (see Figure 6) and production of their statistics reusing and repurposing harmonized components GSIM will enable more flexible and modular production systems. Production will be based upon more standardized applications that are more robust to change and less vulnerable to changing of IT personnel. An increase in the use of standardized applications, which can easily be shared across domains, will enable the IT specialists to more easily work in different domains.

41. The use of GSIM will reduce the workload as many components can be repurposed and reused. This means less repetitive work and more time for innovation.

42. This will free the IT staff to make more robust applications and explore new ways to better meet the changing needs of the statistical organization and their clients at large. This will include more time for creation of robust, modular, harmonized, well documented processes that comply with the requirements of the Common Statistical Production Architecture.

Figure 6.
Design your own imputation process


Impute – GSBPM 5.4
Process Input
Concepts
Structures
Identify potential errors and gaps
This Process Step produces a process metric, "n", the count of potential errors & gaps identified
n = 0
n > 0
Decision point based on count n
Determine appropriate treatment and apply to potential errors and gaps.
Process Output
Concepts
Structures
The Process Input from Business Group records which information objects from the Concepts and Structures Groups are input to the Process Step
The Process Output from Business Group records which information objects from the Concepts and Structures Groups are outputs from the Process Step
Process Step from Business Group
Concepts information objects
Process Control from Business Group
Structures information objects

43. At the international level there will be increased possibilities for co-design and co-development of common components based upon more robust user-requirements from a wider user-community. The IT developers will also have access to a larger development community that all speak the same language to describe their statistical information.

VII. Statistical Data and Metadata Exchange standard, Data Documentation Initiative and other standards

44. As a reference framework of information objects, GSIM has a complementary relationship with standards, such as Statistical Data and Metadata eXchange (SDMX) and Data Documentation Initiative (DDI), which are commonly used to represent and exchange statistical data and metadata.

45. The information objects within GSIM are conceptual; no specific physical representation of the information is prescribed. As a simplified illustration, the name of an organization can be defined as the same concept regardless of whether the information is recorded in a database, in a spread sheet, in a Comma Separated Value(s) (CSV) file, in an Extended mark-up language (XML) file or handwritten on a piece of paper.

46. GSIM allows organizations to start with a common language related to the data and metadata used throughout the statistical production process. In this context, GSIM information objects have been mapped to relevant representations in SDMX and DDI.

47. This will help statistical organizations to describe and manage statistical information using a common language while, at a systems level, the information is represented and exchanged in an appropriate and standard technical format.

48. While GSIM information objects can be mapped to SDMX and DDI (and substantial business benefit can be obtained from harnessing these standards), GSIM does not require these standards to be used. Some producers and some users of statistics may decide to use alternative standards for particular purposes. In other cases, producers of statistics may be open to using SDMX and/or DDI but have legacy information systems which are not economical to update for use with these standards.

49. Describing statistical information using GSIM as the common point of reference helps users identify the relationship between two sets of statistical information which are represented differently from a technical perspective.

50. For example, a statistician may receive some data described in DDI and some described in a locally created format. The statistician can relate both of these to GSIM. The statistician will be able identify which differences are purely technical and which reflect underlying conceptual differences.

51. Once the nature and extent of the differences can be understood, it often proves straightforward to transform the information into a common technical representation (for example, SDMX or DDI) which allows the content to be integrated and explored. This approach ensures that the results of the technical conversion to a common standard are accurately understood, and are sound, from a conceptual perspective.

52. There are a number of synergies between use of GSIM as a reference framework and the application of representation standards such as SDMX and DDI. These synergies have been maximised by design.

53. For example, when determining the set of definitions to be used for information objects within GSIM, existing standards and models were harnessed as key reference sources. While none of these existing sources had the same purpose and scope as GSIM – that is a reference framework of information objects spanning the full statistical production

process – the development of each entailed analysing and supporting particular needs and scenarios related to particular types of statistical data and metadata.

54. In this way GSIM benefited from the investment of time in analysis, modelling, testing and refinement when developing these standards and models to their current level of maturity. It also means GSIM does not vary "for no reason" from terms and definitions which are used in existing standards and models. Where it does vary it is for reasons such as existing relevant standards and models being inconsistent internally, with one another and/or statisticians reporting that alternative terms or definitions are more relevant to their business needs. A direct consequence of this was the revision of the Neuchâtel Model for Classifications, to fully align and integrate it with GSIM.

Annex

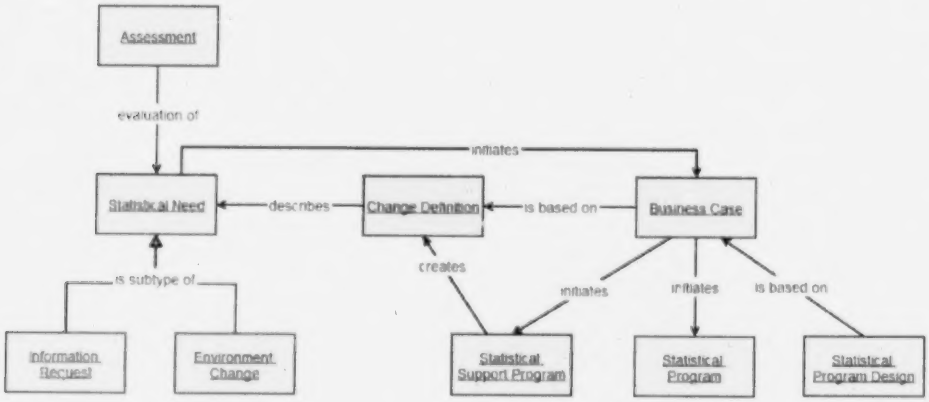
Excerpts from Generic Statistical Information Model Specification document[7]

A. Information in the Statistical Business Process

1. This section looks at different ways that information objects are used within the statistical business process. It considers eight different scenarios, identifying the information objects used and the relationships between those objects.

1. Identifying and Evaluating Statistical Needs

Figure 1.
Identify and Evaluate Statistical Needs



2. An organization will react and change due to a variety of needs. A Statistical Need presents itself to the statistical organization in the form of an Environment Change or an Information Request.

3. Environment Change indicates that there needs to be an externally motivated change. This may be specific to the organization in the form of reduced budget or new demands from stakeholders, or may be a broader change such as the availability of new methodology or technology.

4. When an organization receives an Information Request this will identify the information that a person or organization in the user community requires for a particular purpose. This community may include users within the organization as well as external to it. For example, a the team responsible for compiling National Accounts may need a new Business Process to be initiated to produce new inputs to their compilation process. This request will commonly be defined in terms of a Subject Field that defines what the user wants to measure. When an Information Request is received it will be discussed and clarified with the user. Once clarified, a search will be done to check if the data already

[7] Note: This appendix is not the full GSIM Specification text, it only includes excerpts of that document.

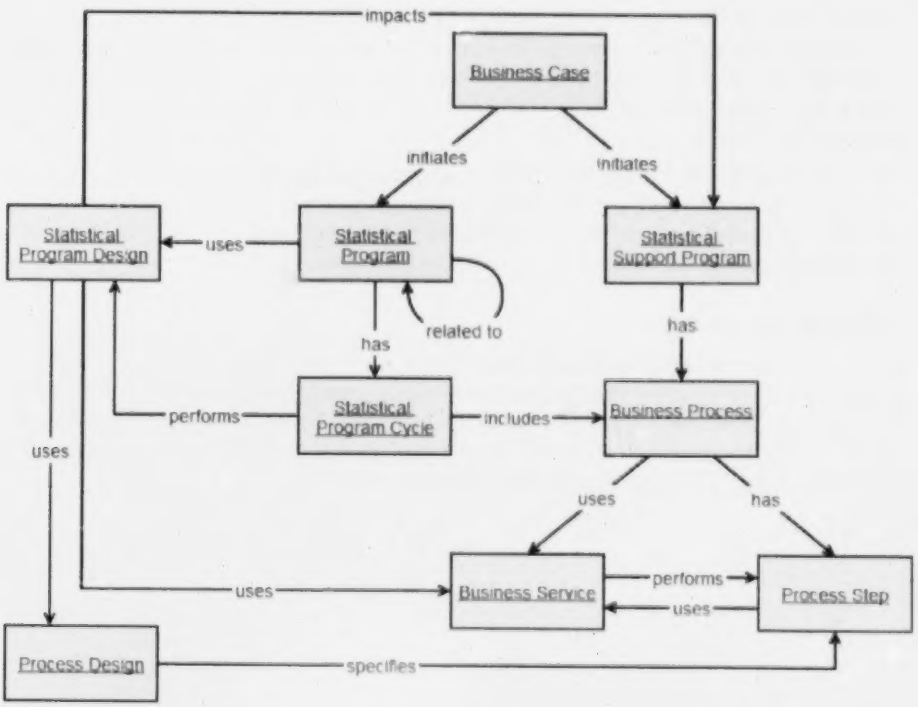
exist. Discovering these Data Sets may be enabled by searching for Concepts and Classifications. Each of these activities are described by a Process Step.

5. The Statistical Need - whether an Information Request or Environmental Change - will be formalized into a Change Definition, typically created by a Statistical Support Programme (a "statistical change programme"). The Change Definition identifies the specific nature of the change in terms of its impacts on the organization or specific Statistical Programmes or Statistical Support Programmes. This Change Definition is used as an input into a Business Case. A successful outcome will either initiate a new Statistical Programme or a new Statistical Support Programme that will create a new Statistical Programme Design that redefines the way an existing Statistical Programme is carried out.

6. A Statistical Need can also be internally driven. At any point in the statistical business process, an organization may undertake an evaluation to determine utility or effectiveness of the business process or its inputs and outputs. An Assessment will be undertaken to evaluate any resources, processes or outputs and may refer to any object described in the model. Assessments include gap analyses undertaken in the context of Business Cases and evaluations undertaken to determine whether a statistical output meets the need for which it was first created.

2. Designing and Managing Statistical Programmes

Figure 2.
Design and Manage Statistical Programme



7. A statistical organization will respond to a perceived Statistical Need by creating a Business Case. Responding to the Business Case will involve one of three things: the creation of a new Statistical Support Programme, the creation of a new Statistical Programme, or the evolution of an existing Statistical Program Design to be implemented by an existing Statistical Programme.

8. Statistical Support Programmes undertake the activities of the statistical organization such as statistical change programmes, data management programmes, metadata management programmes, methodological research programmes, etc. A good example is a programme which manages classifications.

9. Statistical Programmes are those programmes that an organization undertakes to produce statistics (for example, a retail trade survey). Statistical Programmes are cyclical - they perform cycles of collection, production and dissemination of products. Each such cycle is represented by a Statistical Programme Cycle object. The Statistical Programme Cycle is a repeating activity to produce statistics at a particular point in time (for example, the retail trade survey for March 2012).

10. Statistical Programmes require Statistical Programme Designs to achieve their objectives. These designs cover the design of all activities to be undertaken, notably at the level of Business Processes. Within a Statistical Programme Cycle, several Business Processes would typically be performed. These can be understood to correspond to the processes and sub-processes found in the Generic Statistical Business Process Model (GSBPM). These Business Processes may be repeated within a cycle. Each iteration can be made up of multiple activities of the same or different types. As an example of this, within a single cycle, the Statistical Programme might perform three iterations of data collection and processing, then analyse the data and disseminate the resulting statistical Products. Each of these activities could be understood to be a separate Business Process.

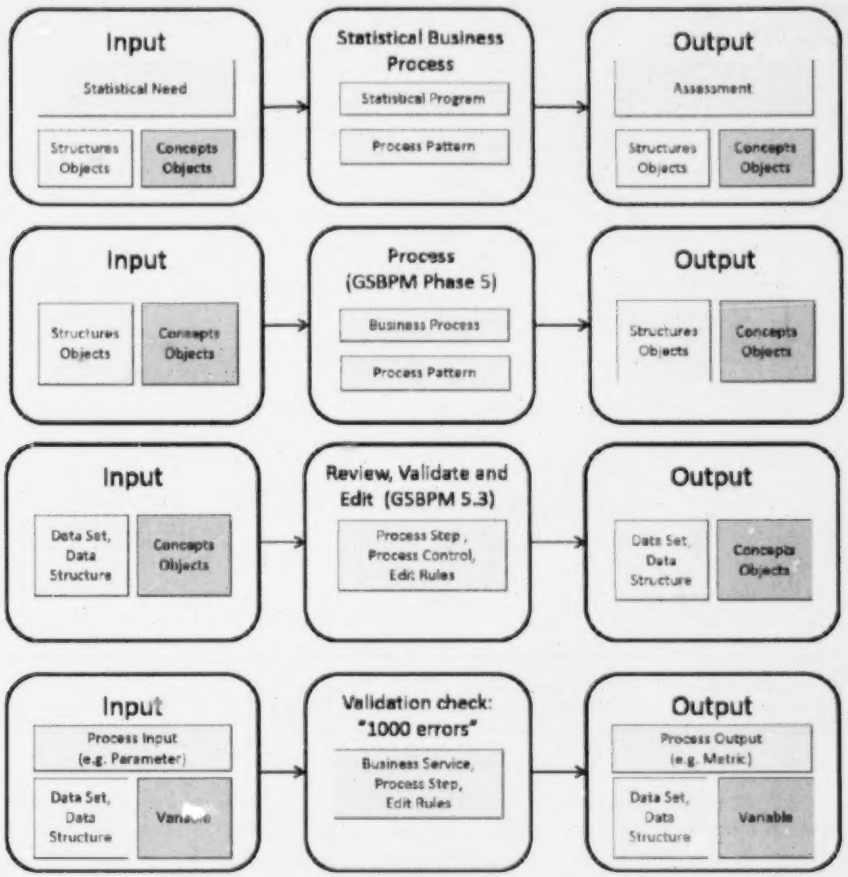
11. The Statistical Programme Design specifies the way in which Business Processes will be conducted. This includes the use of re-usable Business Services (possibly sourced from outside the statistical organization), or through the design and use of more traditional processes. In the latter case, Process Design objects would be used to specify Process Steps. (Although re-usable Business Services are also specified by Process Designs and Process Steps, these will already exist, and not need new design work as part of the Statistical Programme Design.)

12. It should be noted that Statistical Programme Designs specify what Process Steps will need Process Designs, and also which Business Services would be used, but do not do the low-level specification of how such Process Steps and Business Services are executed. These specifications are found in the Process Design object.

3. Designing Process Steps

13. Before explaining the objects which GSIM uses to represent the design of Process Steps, it is important to discuss the nature of processes more generally. The types of objects provided by GSIM perform specific functions. In GSIM, Business Processes have Process Steps. Each Process Step can be as "large scale" or "small scale" as the designer of a particular Business Process chooses (see Figure below).

Figure 3.
Process Steps can be as large or small as needed

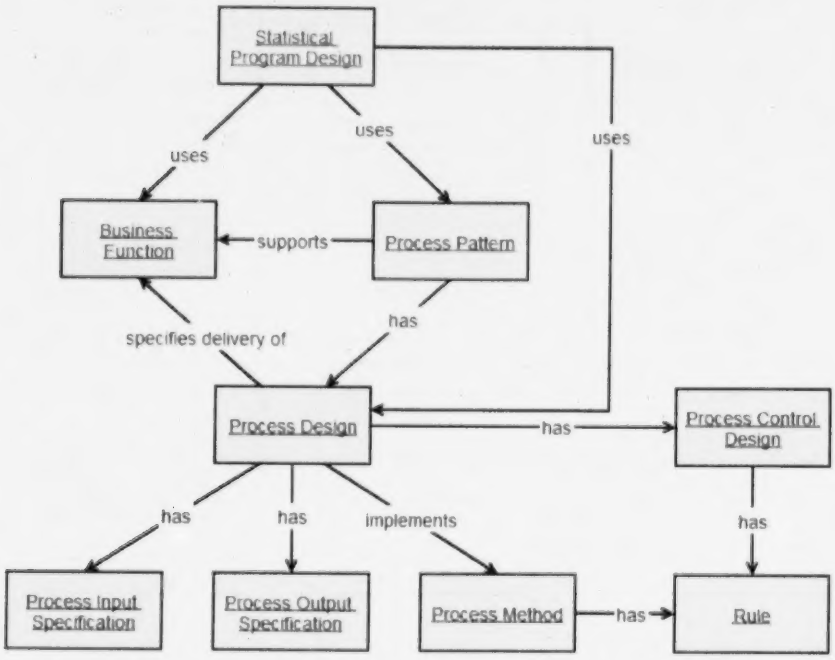


14. Process Steps can contain "sub-steps", those "sub-steps" can contain further "sub-steps" within them and so on indefinitely. Typically, the outputs of one Process Step become inputs to the next Process Step. There can also be conditional flow logic applied to the sequence of Process Steps, based on parameters which have been passed in, or conditions met by the outputs of a previous Process Step.

15. The design of a Process Step thus can be understood to use other Process Steps and even other Business Services which have already been designed and made available for re-use. In a more traditional scenario, the Process Step is designed and then executed. In future, it is foreseen that re-usable Business Services will be increasingly common, having been designed and implemented by another external organization. The next sections describe these two scenarios.

A. Designing Process Steps

Figure 4.
Design Process Steps



16. A Statistical Programme Design is associated with a top level Process Step whose Process Design contains all the sub-steps and process flows required to put that Statistical Programme into effect. Each Process Step in a statistical Business Process has been included to serve some purpose. This is captured as the Business Function associated with the Process Step. An example of a Business Function could be "impute missing values in the data". In order to support this Business Function, an imputation process is needed, which will require a Process Design.

17. In line with the GSIM design principle of separating design and production, GSIM assumes that Process Steps will be designed during a design phase. Having divided a planned statistical Business Process into Process Steps, the next requirement is to specify a Process Design for each step. The Process Design identifies how each Process Step will be performed. A Process Design may use a Process Pattern which is a nominated set of Process Designs and associated flows (Process Control Designs) which have been highlighted for reuse.

18. Process Designs specify several things: they identify the different types of inputs and outputs represented by the Process Input Specification and Process Output Specification. Examples of Process Inputs include data, metadata such as Statistical Classifications, imputation and editing Rules, parameters, etc. Process Outputs can be reports of various types (processing metrics, reports about data validation and quality, etc.), edited Data Sets, new Data Sets, new or revised instances of metadata, etc.

19. To continue the example, the process designer would specify the inputs in the Process Input Specification as imputation Rules and the Data Set for which imputation is desired. The Process Output Specification would include an edited Data Set containing the imputed values, plus a report detailing which values had been imputed.

20. The Process Design specifies the control logic, that is the sequencing and conditional flow logic among different sub-processes (Process Steps). This flow is described in the Process Control Design. When creating a Process Design, a Process Control Design that provides information on "what should happen next" is specified. Sometimes one Process Step will be followed by the same step under all circumstances. In such cases the Process Control Design simply records what Process Step comes next. However, sometimes there will be a choice of which Process Step will be executed next. In this case, the Process Control Design will detail the set of possible "next steps" and the criteria to be applied in order to identify which Process Step(s) should be performed next.

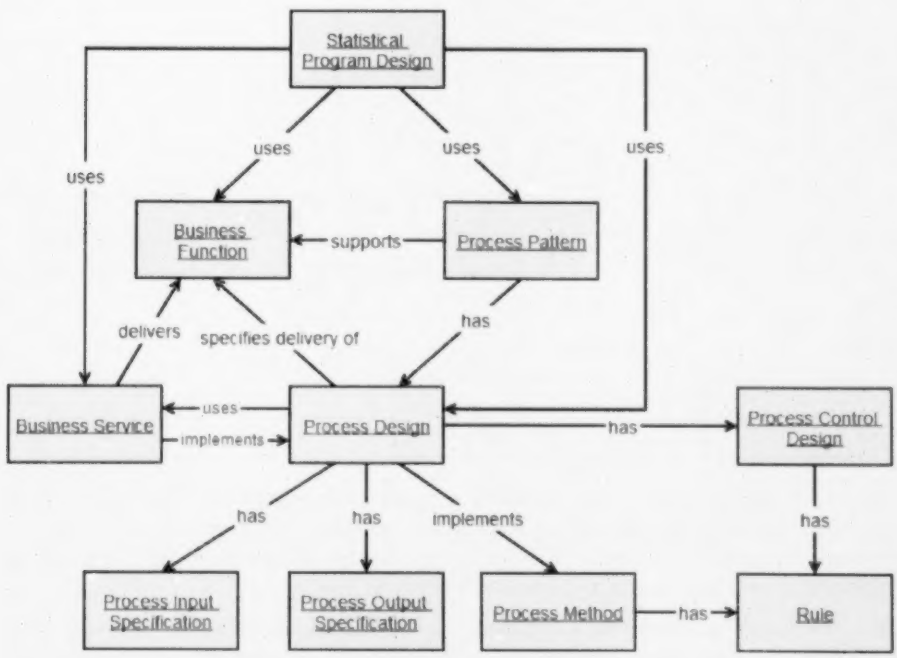
21. The Process Design associated with that Process Step will identify the Process Method that will be used to perform the Business Function associated with the Process Step. For example, if the Business Function is 'impute missing values in the data', the Process Method might be 'nearest neighbour imputation'.

22. A Process Method specifies the method to be used, and is associated with a set of Rules to be applied. For example, any use of the Process Method 'nearest neighbour imputation' will be associated with a (parameterized) Rule for determining the 'nearest neighbour'. In that example the Rule will be mathematical (for example, based on a formula). Rules can also be logical (for example, if Condition 1 is 'false' and Condition 2 is 'false' then set the 'requires imputation' flag to 'true', else set the 'requires imputation flag' to 'false').

23. The resulting Process Design and Process Control Design objects (along with related Process Input Specifications and Process Output Specifications) would be used in the implementation of the Process Step.

B. Using Re-usable Business Services

Figure 5.
Use of re-usable Business Services



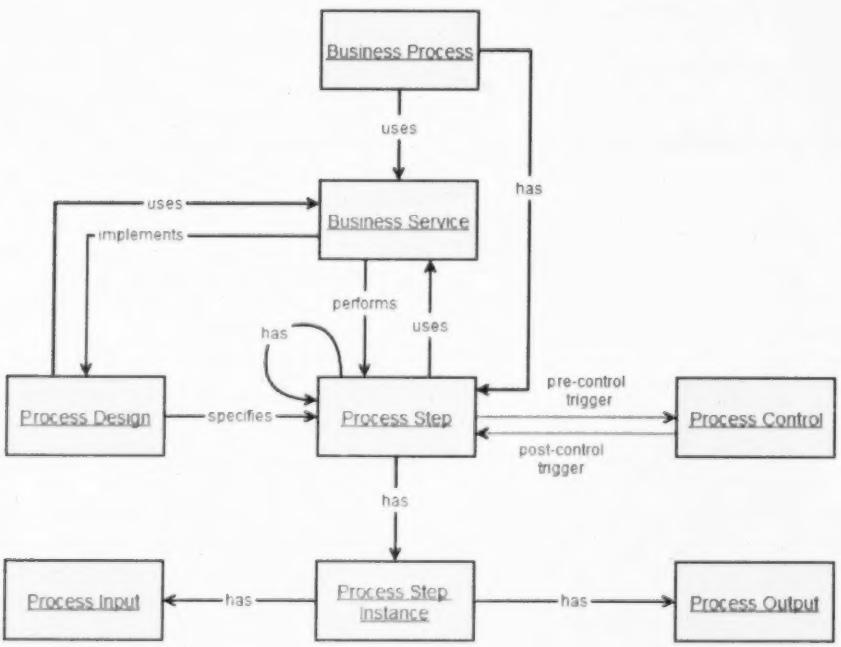
24. It is not always necessary for the Statistical Programme to design its own Process Steps from the beginning. The Common Statistical Production Architecture (CSPA) describes how statistical organizations can create statistical services that are easily reused in other statistical organizations. In GSIM terms, a statistical service is a Business Service. A Business Service is a means of performing a Business Function (an ability that an organization possesses, typically expressed in general and high level terms and requiring a combination of organization, people, processes and technology to achieve).

25. The increased sharing and reuse of Business Services means that the resources needed to meet new demands for statistical production could be considerably reduced, and the time needed to produce new statistical products could be lessened. To facilitate this, CSPA introduced the concept of a statistical services catalogue, where different statistical organizations could list the statistical services they have developed, with the intent of sharing them with other statistical organizations.

26. Business Services have already been designed, with all of the normal input types, output types, process control design, and other properties already specified. Thus, a Business Service can act in a fashion similar to a Process Step designed within the organization, but without the effort required in the traditional scenario.

4. Running Processes

Figure 6.
Run Process



27. A Statistical Programme needs to execute processes to realize some Business Functions. This can be done in two ways: a Process Step can be directly executed by a Business Process, or a re-usable Business Service can be used by the Business Process, as an intermediate trigger for the execution of the Process Step.

28. In order to understand how this works, we characterize the nature of Process Steps in more detail. Process Steps are the resources which have been specified in a Process Design, and which can be executed multiple times. Process Steps can exist at many levels of granularity, and can involve the use of other Process Steps as sub-processes. The navigation

among the sub-processes is performed during execution as indicated by a Process Control, which is itself an implementation of a Process Control Design.

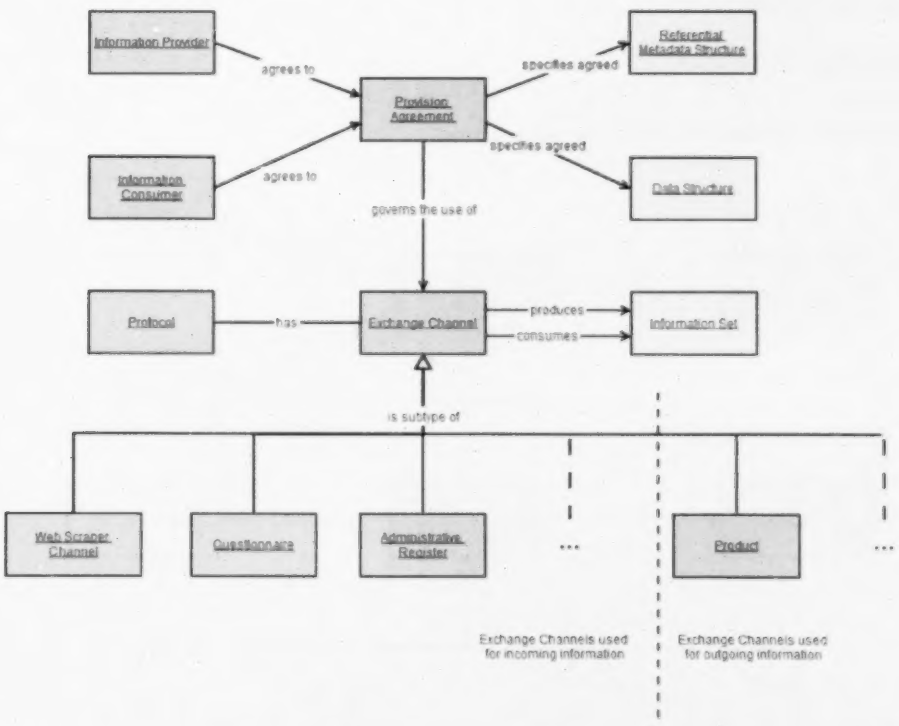
29. Individual executions of a Process Step are represented by the Process Step Instance. It is at this level that specific instances of the inputs and outputs are used. In the Process Design, the types of inputs and outputs are specified (Process Input Specification and Process Output Specification) - the actual instances of inputs and outputs are associated with the Process Step Instance, and are represented by the Process Input and Process Output objects. Inputs can be of any type of information - rules, parameters, data sets, metadata of many kinds, etc. Outputs are similarly of many different types, and often include process metrics and various types of reports, as well as data and metadata.

30. At the time the Process Design is executed someone or something needs to apply the designated method and rules. The Process Design can designate the Business Service that will implement the Process Method at the time of execution. A Business Service represents a service delivered by a piece of software (as described in the section above) or a person. Putting a publication on the statistical institute's website or putting collected response forms in a shared data source for further processing are both examples of Business Services.

31. It should be noted that this model supports both automated and manual processes, and processes which might involve sub-processes of either type.

D. Exchanging Information

Figure 7.
Exchange Channels



32. Statistics organizations collect data and referential metadata from Information Providers, such as survey respondents and providers of Administrative Registers, and disseminate data to Information Consumers, such as government agencies, businesses and

members of the public. Each of these exchanges of data and referential metadata uses an Exchange Channel, which describes the means to receive (data collection) or send (dissemination) information. Information Providers and Information Consumers can be Organizations or Individuals who are either within or external to the statistical organization.

33. Different Exchange Channels are used for collection and dissemination. Examples of collection Exchange Channels include Questionnaire, Web Scraper Channel and Administrative Register. The only example of a dissemination Exchange Channel currently contained in GSIM is Product. Additional Exchange Channels can be added by organizations depending on their needs.

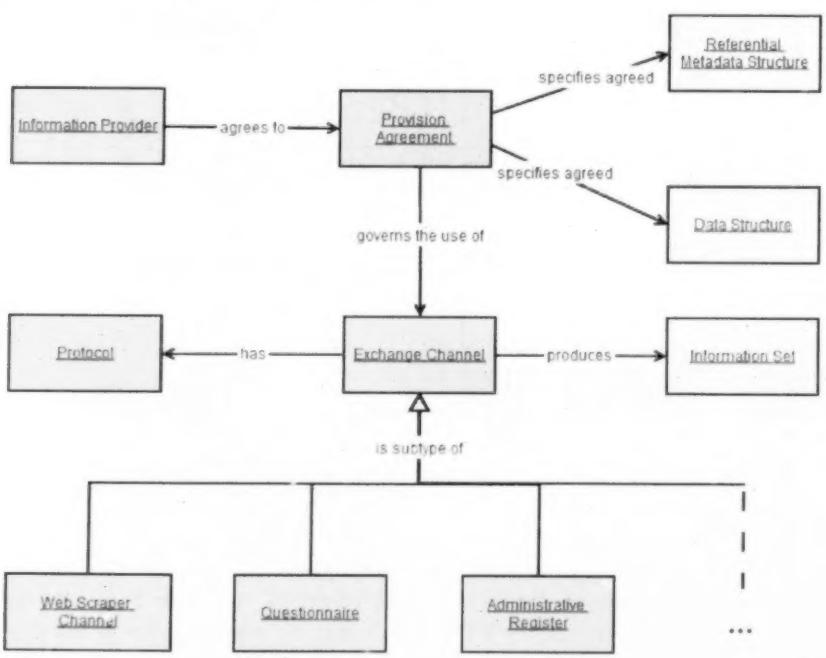
34. The use of an Exchange Channel is governed by a Provision Agreement between the statistics office and the Information Provider (collection) or the Information Consumer (dissemination). The Provision Agreement, which may be explicitly or implicitly agreed, provides the legal or other basis by which the two parties agree to exchange data. The parties also use the Provision Agreement to agree the Data Structure and Referential Metadata Structure of the information to be exchanged.

35. The mechanism for exchanging information through an Exchange Channel is specified by a Protocol (e.g. SDMX web service, data file exchange, face to face interview).

36. To collect data, a statistical organization receives data and referential metadata from the Information Provider in a manner consistent with the Protocol and the Provision Agreement, and the Exchange Channel produces an Information Set. To disseminate data, the Exchange Channel consumes an Information Set, which is then provided to the Information Consumer in a manner consistent with the Protocol and the Provision Agreement. More information about collection and dissemination can be found in the following sections.

5. Collecting Information

Figure 8.
Exchange Channels for collecting information



37. GSIM models three collection Exchange Channel examples: Questionnaire, Web Scraper Channel and Administrative Register. Each of these is detailed in Annex A. Statistics organizations may collect data and referential metadata from Information Providers using additional Exchange Channels, such as file transfer, web services and data scanning. Statistical organizations can extend GSIM to add channels relevant to their context.

38. The use of an Exchange Channel for collection is governed by a Provision Agreement between the statistical organization and the Information Provider. The two parties use the Provision Agreement to agree the Data Structure and Referential Metadata Structure of the data to be exchanged. The mechanism for collecting information through the Exchange Channel is specified by a Protocol (e.g. face to face interview, data file exchange, web robot). The collecting organization uses the collected information to produce an Information Set, which may contain data or referential metadata.

7. Processing and Analysing Information

39. GSIM is very flexible in describing the processing and analysis of information.

40. One can understand the statistical production process from a data-centric perspective[8]. Statistical organizations strive to produce high-quality accurate data that is supported by the metadata needed to make the data optimally useful. For this reason, it is appropriate to think of the evolution of data as it passes through the production process. The focus of many activities is driven by the metadata, but at the end of the production process the metadata is a supporting resource from the perspective of the data and ultimately a statistical product. The relationship of the data and metadata is one which is important to understand.

41. Collected data comes into a statistical organization through an Exchange Channel. Regardless of how the data is collected and where it comes from, it is a resource which will begin a process of evolution through many different stages. The initial data is described as a Data Set with relevant Data Structures. Data Sets are stored in an organised way in a Data Resource. The Data Sets are the primary inputs and outputs of a set of Process Steps, as conducted by a Statistical Program.

42. As the statistical organization moves from raw input data to an increasingly refined set of data, it can be understood that each phase of this processing adds additional Datasets to the Data Resource. There are many different Process Methods which may inform these activities. These are implemented through the different Process Steps that the statistical organization undertakes.

43. At a certain point (and this can take place at different places within the production process, depending on the type of edits being performed) the data will be analysed for the production of statistical Products. The analysis of the data can be understood as using Data Sets from the Data Resource as inputs to processes such as confidentiality routines or to produce explanations of the data. The operations performed during analysis will vary based on what the ultimate Products are - confidentialised unit-record data may be a Product, or we may be publishing aggregated indicators and tables to address specific policy issues, and these involve different types of analysis - but the process is still one of further evolving the information held in the Data Resource.

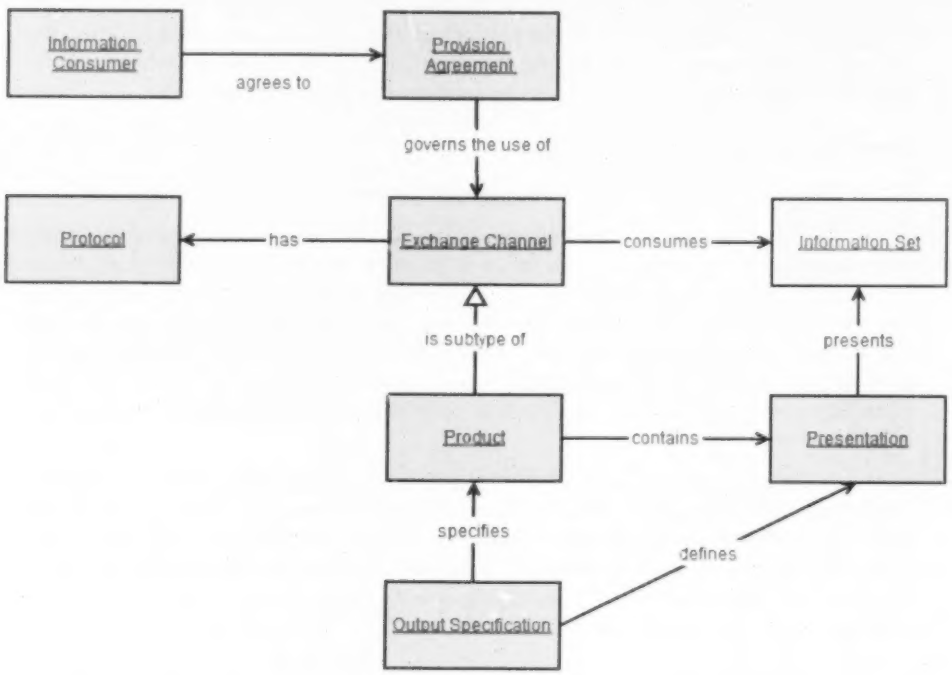
44. In the past, there was an assumption that a data collection will be followed by processing, analysis, and dissemination of the statistical Products. This is a time-consuming and resource-intensive process. One way to make the functions of a statistical organization

[8] Although these paragraphs focus of data, the same descriptions can be applied to referential metadata.

more efficient is to re-use data to produce new Products as they are asked for, lowering the cost and shortening the time needed for production. In this sense, the Data Resource can be understood as an organizational asset, to be managed and exploited to the greatest extent possible.

8. Disseminating Information

Figure 9.
Exchange Channel for disseminating information



45. A statistical organization disseminates statistical information to an Information Consumer.

46. The Information Consumer accesses a set of information via a Product (or potentially via another Exchange Channel), which contains one or more Presentations. Each Presentation will typically provide a view of data and associated metadata to define and describe the structure of the presented data, and perhaps referential metadata in the form of textual media, such as quality reports.

47. A Presentation can take different forms - for example, it could be a screen visualization of a table of data in graphical form displayed in an HTML page, a downloadable PDF, or an SDMX file in XML format.

48. An Output Specification defines what is contained in the Presentation. A Product, which packages Presentations, may be a statistical organization's standard specific output as one might see in:

- a regular statistical bulletin (e.g. a monthly publication of the Retail Prices Index),
- a dynamically generated package of statistical content which is generated following the receipt of a query from an Information Consumer who wishes to access the organization's data via a published API (Application Programming Interface) or

- some data exploration facility which might be built into the statistical organization's website.

49. The Output Specification also defines the information required from the Information Set for the Presentation. The specifications are frequently determined by an internal (to the organization) process which would have specifically standard, static outputs to produce (such as the aforementioned statistical bulletins). For dynamically delivered products, aspects of the specification could be determined by the Information Consumer at run time, via machine to machine dynamic, as exemplified in the API scenario above. In either case, the requests would result in the Output Specification specifying Information Set data and/or referential metadata that will be included in each Presentation.

50. The mechanism for providing a Product is specified by a Protocol (e.g. SDMX-ML, DDI XML, PDF etc.). This formatting information forms part of the Output Specification to generate the Product and its Presentations in the appropriate format.

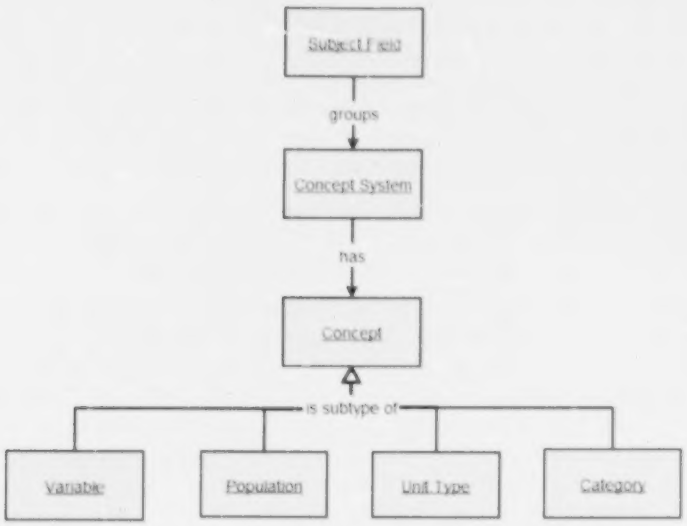
51. The Information Consumer can be one of many forms depending upon the scenario of the request. The Information Consumer could be a person accessing the statistical organization's website and visually inspecting the contents of a web page, or it could be a computer program requesting the information via an API using an SDMX query. The Information Consumer's access to the information would be subject to a Provision Agreement, which would set out the conditions of access and use. This might be in the form of passive acceptance of the terms and conditions of use of the data from a website the Information Consumer is accessing, or in the case of access to a greater level of detail via an API, it might be a more involved registration process.

B. Foundational Information

52. The GSIM Concepts and Structures groups include information objects which are foundational to the statistical Business Process. That is, these objects are the conceptual and structural objects which are used as the Process Inputs and Process Outputs to the process. The Concepts area of GSIM includes the sets of information objects that describe and define the terms used when talking about the real-world phenomena that the statistics measure in their practical implementation. The Structures area includes the set of information objects used in relation to data and referential metadata and their structures. The objects described in this section of the document are used to provide information that helps users of data and metadata understand the results of Business Processes and Statistical Programmes.

1. Concepts

Figure 10.
Concepts



53. At an abstract level, a Concept is defined in GSIM as a 'unit of thought differentiated by characteristics'. Concepts are used in different ways throughout the statistical lifecycle, and each different role of a Concept is described using a different information object (the objects are subtypes of Concept). A Concept can be used in these situations:

(a) As a characteristic. The Concept is used by a Variable to describe the particular characteristic that is to be measured about a Population. For example, to measure the Concept of gender in a population of adults in the Netherlands, the Variable combines this Concept with the Unit Type person.

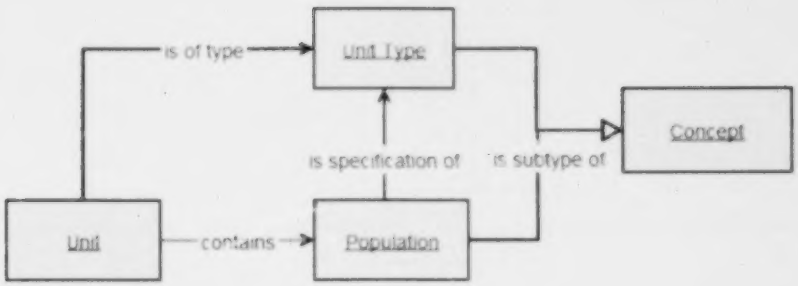
(b) As a Unit Type or a Population. To describe the set of objects that information is to be obtained about in a statistical survey. For example, the Population of adults in the Netherlands, based on the Unit Type of persons.

(c) As a Category to further define details about a Concept. For example, Male and Female for the Concept of Gender. Codes can be linked to a Category via a Node (i.e., a Code Item or Classification Item), for use within a Code List or Statistical Classification.

54. Concept Systems are sets of Concepts which are structured by the relations between those Concepts. A Subject Field groups Concept Systems on the basis of their field of special knowledge (for example, labour market, tourism).

2. Population

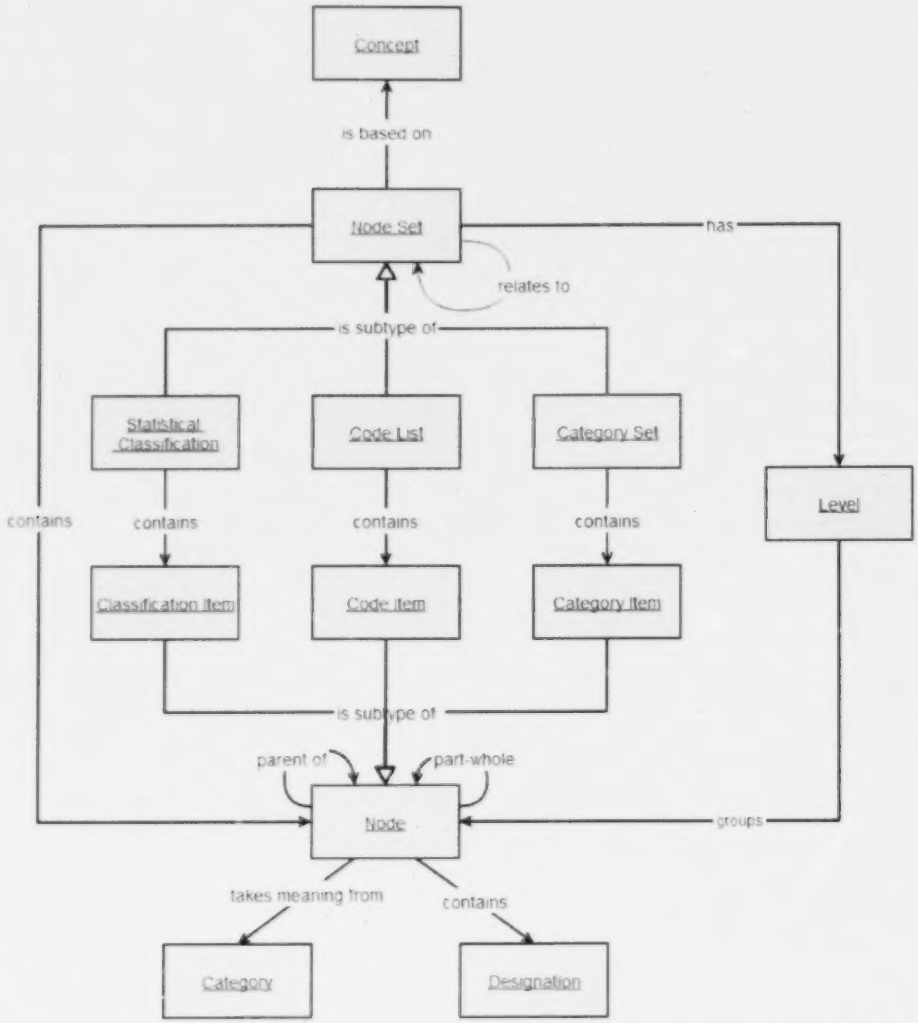
Figure 11.
Populations and Units



55. There are several kinds of Populations depending on what Process Step it is used in. For example a statistical organization may refer to a target, survey, frame, or analysis population. The objects of interest in a statistical process are Units (for example, a particular person or a business). Data are collected about Units. There are two ways in which a unit is specified in the model. A Unit is an individual entity associated with a Population about which information may be obtained. A Unit Type (for example persons or businesses) is a way of identifying an abstract type of Unit that a Variable is measuring.

3. Node and Node Set

Figure 12.
Node and Node Set inheritance



56. A Category is a particular type of Concept whose role is to define a characteristic. There are three ways in which a Category can be used. In GSIM, these are described as the three subtypes of Node - Category Item, Code Item and Classification Item. Categories are grouped into Node Sets based on the way in which it can be used. There are three subtypes of these groups (Node Sets) - Category Sets, Code Lists and Statistical Classifications.

57. A Category Set is a set of Category Items, which contain the meaning of a Category without any associated representations. An example of a Category Set is: Male, Female.

58. In a Code List, the Code Items contain the meaning of the Categories combined with a Code representation. An example of a Code List is: 1. Male, 2. Female.

59. A Statistical Classification is similar to a Code List. It combines the meaning of the Category with a Code representation. However the content of a Statistical Classification must fulfil certain criteria and have a certain status. The Classification Items must be

mutually exclusive and jointly exhaustive for the Level at which they exist at in the Statistical Classification. An example of a Statistical Classification is: 1. Male, 2. Female, 3. Intersex.

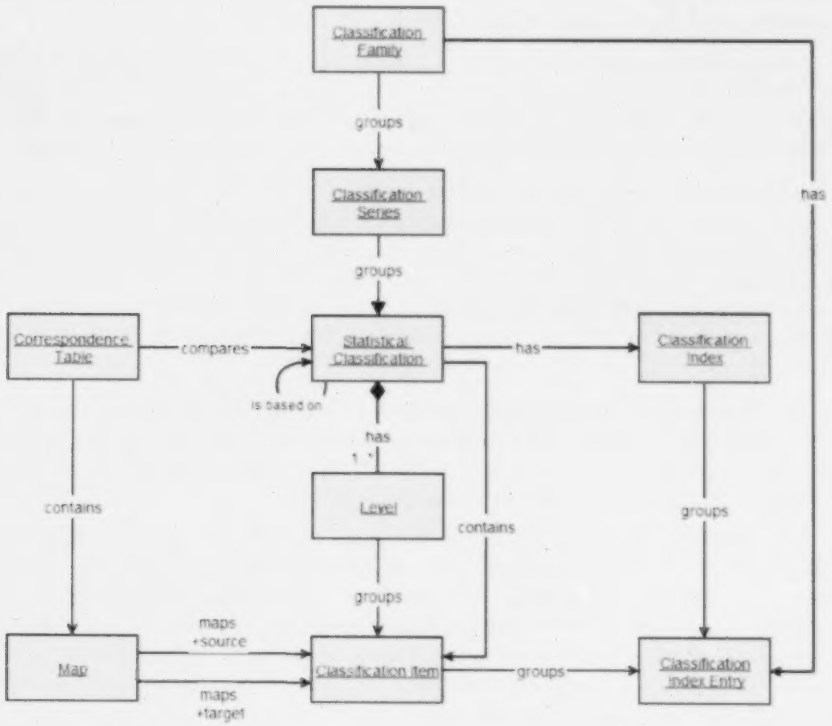
60. A Code List does not have to satisfy the same criteria as the Statistical Classification. The Code List can also contain additional Code Items to support a particular use of the Code List, such as the inclusion of missing values.

61. The similarities between Statistical Classifications, Code Lists and Category Sets are inherent through their link (as subtypes) to Node Set. Similarly, the three types of item which make up each group (Classification Item, Code Item and Category Item respectively) are subtypes of Node.

4. Statistical Classification

62. This section describes a Statistical Classification and its related management objects, as a particular view of the Node Set portion of GSIM. Further detail about Statistical Classifications in particular can be found in the GSIM Statistical Classification Model.

Figure 13.
Statistical Classifications



63. The figure above provides an overview of the objects relating to Statistical Classifications.

64. A Classification Family is a group of Classification Series related based on a common Concept (e.g. economic activity). A Classification Series is an ensemble of one or more Statistical Classifications that are based on the same Concept. The Statistical Classifications in a Classification Series are related to each other as versions or updates. Typically, these Statistical Classifications have the same name, for example International

Standard Industrial Classification of All Economic Activities (ISIC), or International Standard Industrial Classification of Occupations (ISCO).

65. A Statistical Classification is a set of Categories which may be assigned to one or more Represented Variables used in the production and dissemination of statistics. The Categories at each Level of the classification structure must be mutually exclusive and jointly exhaustive of all objects/units in the population of interest. One example of a Statistical Classification is ISIC rev 4.

66. The Categories are defined to reference one or more characteristics of a particular population of interest. A Statistical Classification may have a flat, linear structure or may be hierarchically structured, such that all Categories at lower Levels are sub-categories of a Category at the next Level up.

67. A Statistical Classification has Categories that are represented by Classification Items. These Classification Items are organised into Levels determined by the hierarchy. A Level is a set of Concepts that are mutually exclusive and jointly exhaustive; for example: section, division, group and class in ISIC rev 4.

68. A Classification Item combines the meaning from a Category, its representation (i.e., Code) and additional information in order to meet the Statistical Classification criteria, for example "A- agriculture, forestry and fishing" and accompanying explanatory text such as information about what is included and excluded.

69. Statistical Classifications can be versions or variants. A variant type of Statistical Classification is based on a version type of Statistical Classification. In a variant the Categories of the version may be split, aggregated or regrouped to provide additions or alternatives to the standard order and structure of the original Statistical Classification.

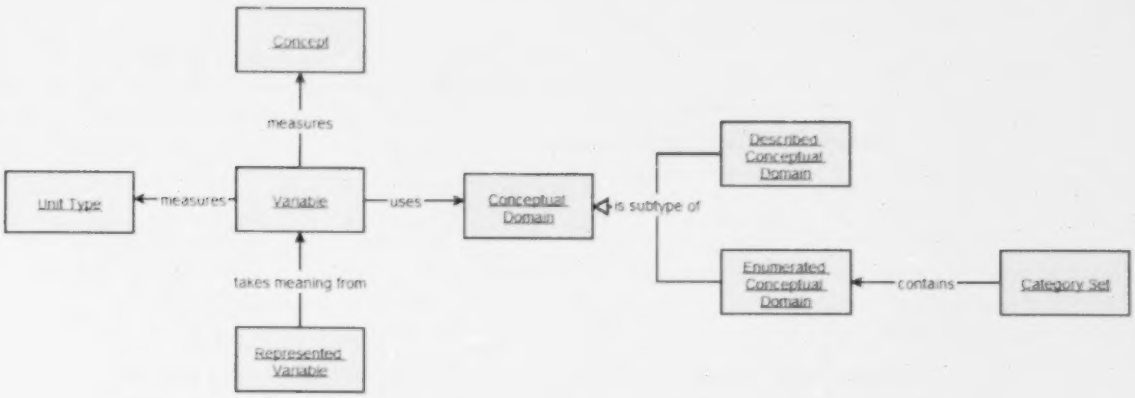
70. A Correspondence Table is a set of Maps. These Maps link a Classification Item in a Statistical Classification with a corresponding Classification Item in another Statistical Classification via the Concept which is common to both Classification Items. For example, in a Correspondence Table displaying the relationship between ISIC rev 4 and the North American Industry Classification System (NAICS 2007 (US)), "0112 - Growing of Rice" in ISIC Rev 4 is related to "111160 - Rice Farming" in NAICS through the common concept of "growing rice".

71. A Classification Index shows the relationship between text found in statistical data sources (responses to survey questionnaires, administrative records) and one or more Statistical Classifications. A Classification Index may be used to assign the Codes for Classification Items to observations in Statistical Programs.

72. A Classification Index Entry is a word or short text (e.g. the name of a locality, an economic activity or an occupational title) describing a type of Concept to which a Classification Item applies, together with the Code of the corresponding Classification Item. Each Classification Index Entry typically refers to one item of the Statistical Classification. Although a Classification Index Entry may be associated with a Classification Item at any Level of a Statistical Classification, they are normally associated with Classification Items at the lowest Level.

5. Variable

Figure 14.
Variable



73. When used as part of a Business Process, a Unit Type defining a Population is associated with a characteristic. The association of Unit Type and a Concept playing the role of a characteristic is called a Variable (see Figure 14). For example, if the Population is adults in Netherlands, then a relevant Variable might be the Concept educational attainment combined with the Unit Type person.

74. The Variable (person's educational attainment) does not include any information on how the resulting value may be represented. This information (the Value Domain) is associated with the Represented Variable. This distinction promotes the reuse of a Variable definition when what is being measured is conceptually the same but it is represented in a different manner.

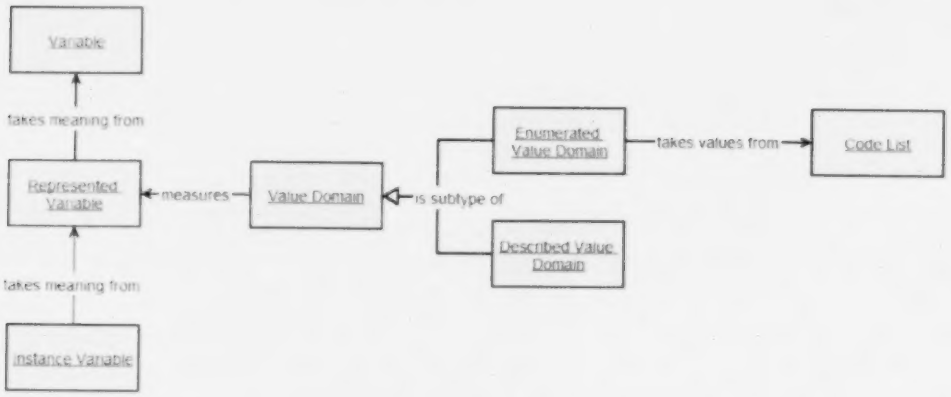
75. A derived variable is created by a Process Step that applies a Process Method to one or more Process Inputs (Variables). The Process Output of the Process Step is the derived variable.

76. A Conceptual Domain is associated with a Variable. It has two subtypes: Described Conceptual Domain and Enumerated Conceptual Domain. An Enumerated Conceptual Domain, in combination with a Category Set, contains information on the semantics of the Categories used by the Variable.

6. Represented Variable

77. GSIM assists users in understanding both the meaning and the concrete data-representation of the object. Accordingly, GSIM distinguishes between conceptual and representation levels in the model, to differentiate between the objects used to conceptually describe information, and those that are representational.

Figure 15.
Represented Variable



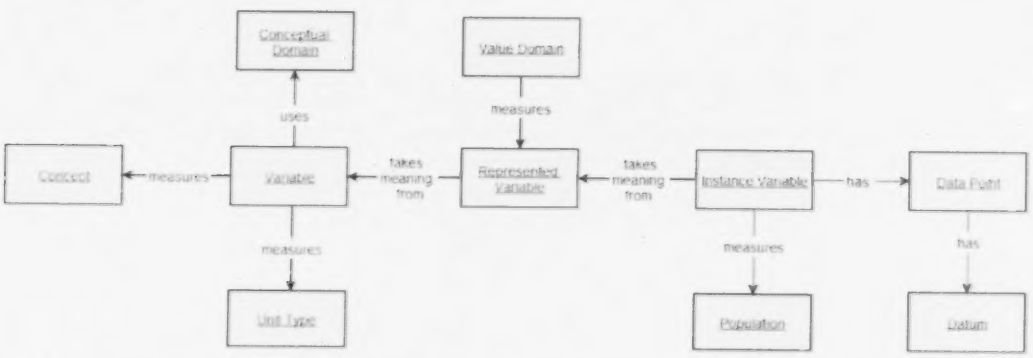
78. The Represented Variable (see Figure 15) adds information that describes how the resulting values may be represented through association with a Value Domain. While Conceptual Domains are associated with a Variable, Value Domains are associated with a Represented Variable. These two domains are distinguished because GSIM separates the semantic aspect (Conceptual Domain) and the representational aspect (Value Domain).

79. Both the Enumerated Value Domain and the Described Value Domain (the two subtypes of Value Domain) give information on how the Represented Variable is represented. The Enumerated Value Domain does this in combination with a Code List, while the Described Value Domain provides a definition of how to form the values, rather than explicitly listing them.

80. The Value Domain includes data type and unit of measure information. The data type contains information on the allowed computations one may perform on the Datum (nominal-, ordinal-, interval-data, etc.), while the unit of measure (Tonnes, Count of __, Dollars, etc.) refines the measure of the Value Domain. For example gender codes lead to nominal statistical data, whereas age values in years lead to interval data.

6. Instance Variable

Figure 16.
Instance Variable



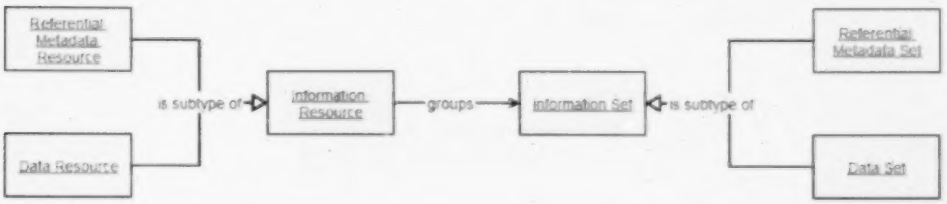
81. An Instance Variable (see Figure 16) is a Represented Variable that has been associated with a Data Set. This can correspond to a column of data in a database. For example, the "age of all the US presidents either now (if they are alive) or the age at their deaths" is a column of data described by an Instance Variable, which is a combination of

the Represented Variable describing "Person's Age" and the Value Domain of "decimal natural numbers (in years)".

82. A Datum is contained within a Data Point in a Data Set. It may be defined by the measure of a Value Domain associated with a describing Instance Variable, combined with the link to a Unit (for unit data), or a Population (for dimensional data).

8. Information Resources

Figure 17.
Information Resources



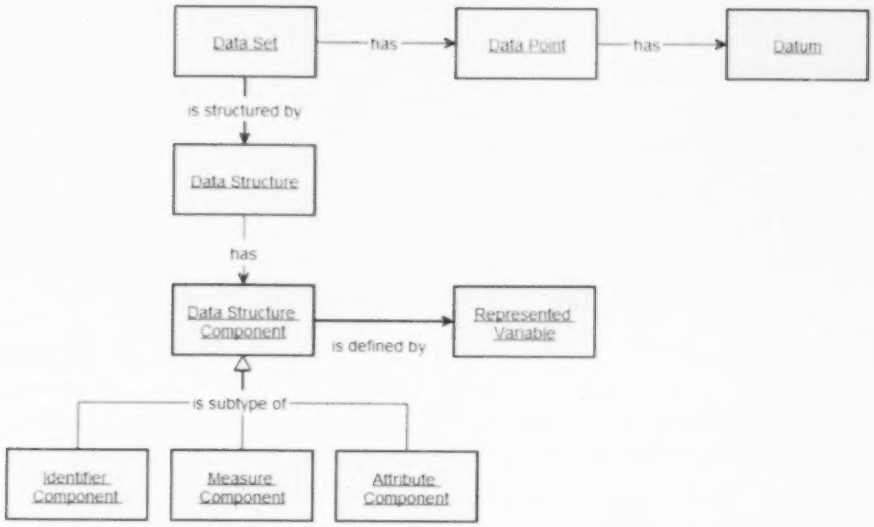
83. Statistical organizations collect, process, analyse and disseminate Information Sets, which are either data (Data Sets) or referential metadata (Referential Metadata Sets).

84. Each Data Set must be structured according to a Data Structure (for example, a structure for Balance of Payments, Demography, Tourism, Education etc.). In the same way, a Referential Metadata Set must be structured according to a Referential Metadata Structure (e.g. an organization's quality framework).

85. Information Resources contain Information Sets. The main purpose of the Information Resource is to aid discovery and management of Information Sets, by providing location and other information relevant to these tasks. There are two types of Information Resource. Data Resources contain Data Sets, and Referential Metadata Resources contain Referential Metadata Sets.

9. Data Sets

Figure 18.
Data Sets



86. A Data Set has Data Points. A Data Point is placeholder (for example, an empty cell in a table) in a Data Set for a Datum. The Datum is the value that populates that placeholder (for example, an item of factual information obtained by measurement or created by a production process). A Data Structure describes the structure of a Data Set by means of Data Structure Components (Identifier Components, Measure Components and Attribute Components). These are all Represented Variables with specific roles.

87. Data Sets come in different forms, for example as Administrative Registers, Time Series, Panel Data, or Survival Data, just to name a few. The type of a Data Set determines the set of specific attributes to be defined, the type of Data Structure required (Unit Data Structure or Dimensional Data Structure), and the methods applicable to the data.

88. For instance, an administrative register is characterized by a Unit Data Structure, with attributes such as its original purpose or the last update date of each record. It contains a record identifying variable, and can be used to define a Population that is used as a frame, to replace or complement existing surveys, or as an auxiliary input to imputation. Record matching is an example of a method specifically relevant for registers.

89. An example for a type of Data Set defined by a Dimensional Data Structure is a time series. It has specific attributes such as frequency and type of temporal aggregation and specific methods, for example, seasonal adjustment, and must contain a temporal variable.

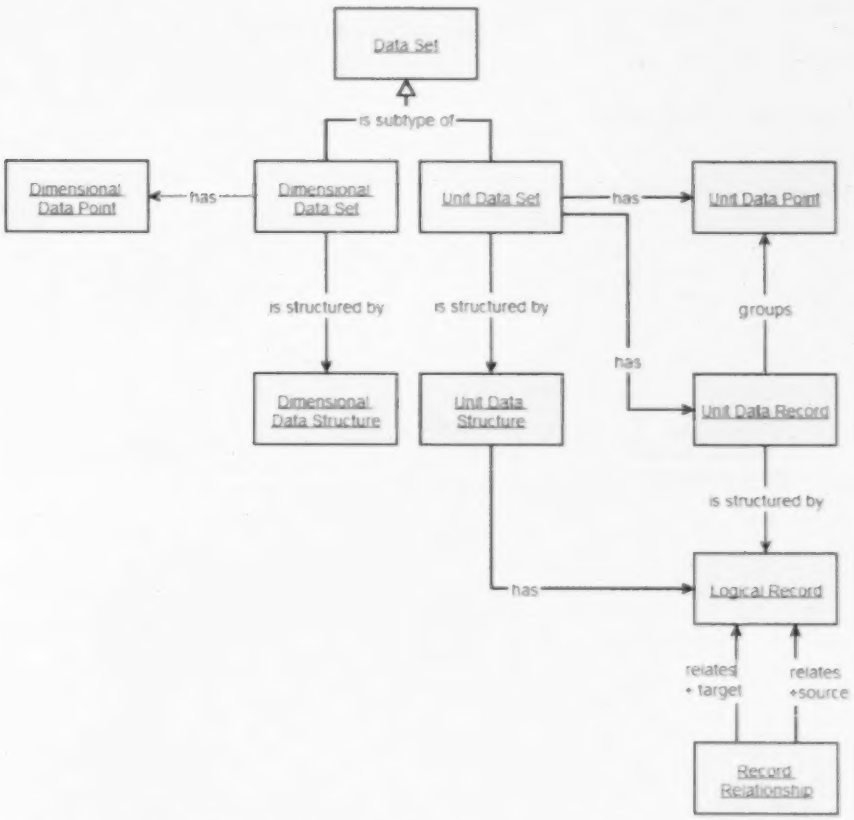
90. Unit data and dimensional data are perspectives on data. Although not typically the case, the same set of data could be described both ways. Sometimes what is considered dimensional data by one organization (for example, a national statistical office) might be considered unit data by another (for example, Eurostat where the unit is the member state). A particular collection of data need not be considered to be intrinsically one or the other. This matter of perspective is conceptual. In GSIM, the distinction is that a Unit Data Set contains data about Units and a Dimensional Data Set contains data about either Units or Populations.

91. GSIM states that all Data Sets must have a structure associated with them. There are, however, cases where a Data Set has no structure – because it was not stored or lost, or it is not known. This type of data may become more prevalent for statistical organizations in the

future. In order for a statistical organization to use this data, the data will need to go through a process of being structured. For example, in a case of investigation of new potential data sources for a new or changed Statistical Need, there will need to be a process where these new data are analysed to determine their content and structure. It is only after this process that these new Data Sets can be described using the Data Structure objects. This unstructured data is currently described by GSIM as a Process Input. Organizations could extend GSIM to capture this use case by creating a new subtype of the Information Set object.

10. Dimensional and Unit Data Structures

Figure 19.
Data Structures



92. A Dimensional Data Structure describes the structure of a Dimensional Data Set by means of Represented Variables with specific roles.

93. The combination of dimensions contained in a Dimensional Data Structure creates a key or identifier of the measured values. For instance, country, indicator, measurement unit, frequency, and time dimensions together identify the cells in a cross-country time series with multiple indicators (for example, gross domestic product, gross domestic debt) measured in different units (for example, various currencies, per cent changes) and at different frequencies (for example, annual, quarterly). The cells in such a multi-dimensional table contain the observation values.

94. A measure is the variable that provides a container for these observation values. It takes its semantics from a subset of the dimensions of the Dimensional Data Structure. In

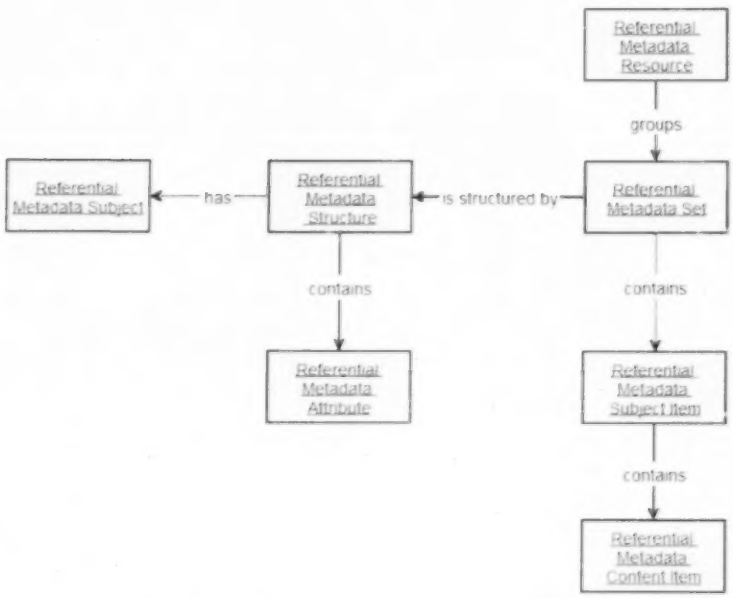
the previous example, indicator and measurement unit can be considered as those semantics-providing dimensions, whereas frequency and time are the temporal dimensions and country the geographic dimension. An example for a measure in addition to the plain 'observation value' could be 'pre-break observation value' in the case of a time series. Dimensions typically refer to Represented Variables with coded Value Domains (Enumerated Value Domains), measures to Represented Variables with uncoded Value Domains (Described Value Domains).

95. A Unit Data Structure describes the structure of a Unit Data Set by means of Represented Variables with specific roles. It distinguishes between the logical and physical structure of a Data Set. A Unit Data Set may contain data on more than one type of Unit, each represented by its own record type.

96. Logical Records describe the structure of such record types, independent of physical features by referring to Represented Variables that may include a unit identification (for example, household number). A Record Relationship defines source-target relations between Logical Records.

11. Referential Metadata Sets

Figure 20.
Referential Metadata Sets



97. Information that describes the characteristics of statistics is "referential metadata". These metadata can be broad, such as about an entire Statistical Program, or narrow, such as about an individual Data Point. Referential Metadata Resources, a special type of Information Resource, provide top-level containers for referential metadata.

98. A Referential Metadata Set organizes referential metadata, whose structure is defined in a Referential Metadata Structure. A Referential Metadata Structure specifies both the Referential Metadata Subject for which referential metadata may be included, and a structured list of Referential Metadata Attributes that can be reported or authored for the given Referential Metadata Subject.

99. These subjects may be any GSIM object type, or any Data Point or set of Data Points created from a specific Data Structure.

- Example of a GSIM object type as a Referential Metadata Subject: Product for which there is a list specified in a Value Domain. The Value Domain specifies the list of actual Products for which reference metadata can be reported or authored using this Referential Metadata Structure.
- Examples of Referential Metadata Attributes include status, coverage, methodology description, and quality indicator.

100. A Referential Metadata Set contains the actual referential metadata reported or authored. The Referential Metadata Subject Item identifies the actual object e.g. actual Product such as Balance of Payments and International Investment Position, Australia, June 2013, or actual Data Points such as the Data Points for a single region within a Data Set covering all regions for a country.

101. The Referential Metadata Content Item is the actual metadata for the identified Referential Metadata Subject Item. Each Referential Metadata Content Item contains the reported referential metadata for one Referential Metadata Attribute specified in the Referential Metadata Structure.

Table 1.
Example of Use of GSIM Referential Metadata Objects

GSIM Object	ONS Statistical bulletin: Public Sector Finances, October 2013:**Table 1**
Referential Metadata Structure	Implicit
Referential Metadata Subject	Data Structure Component
Referential Metadata Attribute	Table footnote
Referential Metadata Set	Footnotes
Referential Metadata Subject Item	*Data Structure Component*: £ billion; PS Current Budget; PS Current Budget ex APF;
Referential Metadata Content Item	Footnoted text